MASTER OF SCIENCE BY RESEARCH

Review of the Evidence on the Use of Arbitration or Consensus within Breast Screening; A Systematic Scoping Review

Hackney, Lisa

Award date:
2016

Awarding institution:
Coventry University

Master of Science by Research in Clinical Practice: Review of the Evidence on the Use of Arbitration or Consensus within Breast Screening; A Systematic Scoping Review.

By Lisa Hackney

May 2016

A thesis submitted in partial fulfilment of the University's requirements for the Degree of Master of Science by Research in Clinical Practice

Abstract

Breast screening involves the interpretation and reporting of mammographic images. In the UK, the images are independently double reported, and inherent with this strategy is that readers may disagree with their decision as to whether a potential abnormality requires further investigation. Discrepant findings require resolution, which is currently achieved by some form of arbitration or consensus.

The primary focus of this scoping review was to establish what evidence is there to inform arbitration and consensus processes and what is their effectiveness within mammography reporting. A systematic scoping review was undertaken to identify the 'nature and extent of research evidence'.

The first stage of the process describes the various sources of information that were searched (databases, conference proceedings, personal contacts and unpublished data sources) using varying search strategies. A 3-stage process was utilised to screen a large volume of literature (601) against the inclusion and exclusion criteria. 26 papers were retained for the final review.

The results of the data extraction were synthesized into key features and emerging themes, with generalizability discussed relative to UK practice. The review has identified a lack of guidance and underpinning evidence to inform how best to use arbitration or consensus to resolve discordant reads. The strengths and weaknesses of the study are discussed and recommendations for future research.

Acknowledgements

The completion of the thesis was made possible by an HEE/NIHR award. The views expressed in this thesis are those of the author.

I am very grateful to my Project supervisors who offered immense support and encouragement:

Project supervisors:

Becky Whiteman – Director of Studies

Professor Ala Szczepura

Dr Louise Moody

Colleagues, friends and family

I would also like to thank my incredible friends and colleagues for encouraging me to undertake the project. In particular, my partner Darren Bould for his unwavering belief in me to achieve my 'goals'.

Contents

Attachment One Clinical Portfolio – submitted separately.

Tables and Figures

Glossary

Arbitration The use of a third reader to decide on case management when there is disagreement between the initial reporters.

Blinded reading The 2nd reader is unaware of the 1st readers report

Consensus A group of film readers who decide on case management when there is disagreement between the initial reporters

Craniocaudal view A standard screening projection of the breast. The x-ray beam enters at the cranial aspect of the breast and exits at the caudal aspect.

Double-reporting A breast screening protocol in which two film readers report the same images independently.

False negative A cancer erroneously discharged at screen reporting or assessment.

False positive A normal case incorrectly recalled for assessment.

Full-field digital mammography The direct acquisition of a digital X-ray image of the breast.

Incident screen A follow-up screening mammogram after a predetermined interval

Interval cancer A cancer that presents clinically between screening rounds.

Medio-lateral oblique view A standard screening projection of the breast taken at an oblique angle

Prevalent screen The first screening mammogram

Reader An individual trained to report breast-screening mammograms

Report The reader's final opinion on a screening mammogram

Abbreviations

2-Dimensional (2D)

3-Dimensional (3D)

ACR American College of Radiologists

BIRADS Breast Imaging Reporting and Data System

BSBR British Society of Breast Radiology

CAD Computer Aided Detection

CASP Critical Appraisal Skills Programme

CBE Clinical Breast Examination

CI Collective Intelligence

CPD Continuing Professional Development

ECR European Congress of Radiology

FFDM Full-Field Digital Mammography

FNAC Fine Needle Aspiration Cytology

MDT Multidisciplinary Team

NA Not Applicable

NBCSP Norwegian Breast Cancer Screening Program

NBSS National Breast Screening Service

NCB Needle Core Biopsy

NCIN National Cancer Information Centre

NCRI National Cancer Research Institute

NDROR Non-Discordant Radiographer Only Reporting

NHS National Health Service

NHSBSP National Health Service Breast Screening Programme

PPV Positive Predictive Value

QA Quality Assurance

RCR Royal College of Radiologists

RCRBG Royal College of Radiologists Breast Group

RSS Rich Site Summary

SBI/ACR Society of Breast Imaging/American College of Radiology Breast Imaging Symposium

SCoR Society and College of Radiographers

SDR Standardised Detection Ratio

U/S Ultrasound

VAB Vacuum Assisted Biopsy

Chapter 1. Introduction

This thesis is set in the context of breast screening, which involves the interpretation and reporting of mammographic images. In the UK, the images are independently double reported, and inherent with this strategy is that readers may disagree with their decision as to whether a potential abnormality requires further investigation. Discrepant findings require resolution, which is currently achieved by some form of arbitration or consensus. The primary focus of this scoping review is to establish what evidence is there to inform arbitration and consensus processes and their effectiveness within mammography reporting.

1.1 Breast Screening Context

1.1.1 Breast Cancer Incidence Worldwide

An estimated 1.6 million women were diagnosed with breast cancer worldwide in 2012, representing the most common cancer in developed and developing countries (Ferlay et al. 2013). The strongest risk factor for female breast cancer is age, and with population ageing, it is therefore predicted that this rate will continue to rise.

1.1.2 Breast Cancer Costs in the UK

As with any healthcare provision, there is the consideration of the costs of screening, diagnosis and treatment. The National Cancer Information Centre (NCIN 2012) report that breast cancer costs in the UK amount to an excess of £5.7 billion annually

(NCIN 2012) with £441.0M projected inpatient costs for 2016 (Local Cancer Intelligence 2016).

1.1.3 UK Breast Cancer Incidence

The most recent data (2013) for which statistics are publically available in the UK, demonstrate that '*breast cancer is the most frequent female cancer'* (Figure 1), with '*1 in 8 women developing the disease in their lifetime*' and nearly '*12,000 subsequent related deaths*' annually (Office for National Statistics 2015). Cancer Registration statistics (2013) validates that 43.5% of UK female breast cancer cases are diagnosed in the 50-59-age range and 34.3% in the 60-69-age range, with a 6% increase in incidence rates in UK females between 2002-2004 and 2011-2013.

Figure 1 Most Common Cancers in Females, UK, 2013. Cancer Research UK,

http://www.cancerresearchuk.org/health-professional/cancer-statistics/incidence/common-cancers-compared#heading-Two, Accessed February 2016.

The combination of breast cancer prevalence and demographic trends contributed to the founding of the UK National Health Service (NHS) Breast Screening Programme (NHSBSP), which was inaugurated in 1988 (Forrest 1986) to facilitate early detection of the disease. The fundamental purpose of the NHS is to improve health and well-being (Department of Health 2008) and although the incidence of breast cancer has continued to rise in the UK over the last decade (Figure 2) the mortality rates from the disease have fallen.

Figure 2 Percentage Change in Female Mortality Rates. Cancer Research UK,

www.cancerresearchuk.org/sites/default/files/cstream-node/cs_mort_20common_females.pdf, Accessed February 2016.

In recent years, some controversy has emerged over the continued overall benefit of breast screening. This led to the commissioning of an independent investigation to evaluate the evidence on '*benefits and harms of breast screening'* from a UK perspective (Marmot et al. 2013: 226). Autier et al. (2010:4411) report that a 35% reduction in breast cancer mortality in England and Wales (period 1989 -2006) can be credited to '*improved treatments (surgical, radiotherapy/chemotherapy regimes and hormonal therapies) and improved delivery of specialist care by multi-disciplinary teams'.*

Overall the Marmot report concluded that screening does '*prevent around 1,300 breast cancer deaths in the UK per year*' (Marmot et al. 2013: 226). However, approximately '*4,000 women each year aged 50 to 70 in the UK'* will undergo '*treatment for a condition that would never have caused them harm'* (Marmot et al. 2013: 226). With this information in mind, it is imperative that screening programmes maximise cancer detection while minimising excessive false-positive recalls and recalls for non-life threatening disease.

1.2 Breast Screen Reporting

Whilst double reporting is the norm in many European countries, the professionals undertaking the task differ in that the UK is unique in utilising non-medical practitioners; radiographers trained to specialise in breast imaging reporting. Double reporting requires that two individuals independently read the mammogram and make a final conclusion (report) on case management. If there is disagreement

between the two reporters there needs to be a resolution on whether to recall for further investigation, or conclude the case is normal and discharge to routine three yearly screening. When a discrepant case is identified, the two reporters may attempt to reach agreement through discussion (consensus), or send to a panel of reporters for evaluation (consensus). Alternatively, an independent third person may review the case and make the final decision (arbitration).

Upon implementation of the NHSBSP in 1988, Consultant Radiologists or other medically qualified individuals were the only ones considered to have the necessary skill set to read and interpret mammographic images. However, severe radiologist shortages necessitated a change in service delivery. Literature began to emerge that Radiographer interpretation of mammograms was as accurate as that of Radiologists (Pauli et al., 1996). The NHS Plan (DH 2000) was pivotal in developing advanced radiographic practice, and radiographer reporting of screening mammography formed part of this strategy (DH 2007a).

In 2011, the DOH recognised that radiographer reporting was crucial to attain health service improvements whilst maximising patient care (DH 2011).

1.3 Changing UK Context

Guidance on arbitration personnel within the UK NHSBSP is currently under review by an expert group. It is recognised that, to maintain the current quality standards and avoid delays in patient management, the extension of arbitration duties to non-medics may now need to be considered (Bennett et al. 2012). Concerns about the

future availability of specialist radiologists have been highlighted in a recent Royal College of Radiologists (RCR 2016) publication. This predicts the retirement of 21% of breast radiologists in the next five years, together with a potential 2.2 million increase in women eligible for screening if the age extension is implemented (based on current population figures). Changing UK practice reflects a shift in focus away from job titles to quality standards for tasks and responsibilities (NHSBSP 2011). It is, therefore, an opportune time to establish what evidence there is to support different models of arbitration or consensus review in breast screening.

1.4 Thesis Aims and Objectives

1.4.1 Aim

The aim of this thesis is to review and evaluate the evidence to support the effectiveness of different strategies utilised to resolve discordance in breast screening reports. For the purpose of this thesis effectiveness is defined in terms of recall rates, cancer detection rate, Positive Predictive Value (PPV) and programme sensitivity/specificity. Although cost-effectiveness is an essential component of service delivery, it is not considered in this thesis.

1.4.2 Objectives

The main objectives of this research are:

Objective 1: To identify the '*nature and extent of research evidence'* (Booth, Papaioannou and Sutton 2012: 27) that has been undertaken on arbitration and

consensus processes within mammography reporting.

Objective 2: To synthesise the known evidence for the effectiveness of the processes used to resolve discordant reports and to identity gaps in the evidence base to inform further research if required.

1.5 Chapter 1 Summary

This chapter has identified the prevalence and trends of breast cancer and the rationale for the founding of the UK NHSBSP. Controversies of the continued overall benefit of breast screening were discussed. The concept of double reading was introduced and that the UK implemented radiographer reporting to sustain this. Double reading inherently results in discordant cases, which require resolution. The next chapter details the NHSBSP processes, reporting strategies and methods to resolve discordant reports.

Chapter 2. Background

The following section describes the complexities of the UK NHSBSP, with differences relevant to other countries highlighted. After clarifying the NHSBSP process, assessment, key standards, and mammography as a screening examination, the chapter introduces the different strategies utilised for reporting.

2.1 UK National Health Service Breast Screening Programme (NHSBSP)

Breast screening primarily entails the mammographic imaging of asymptomatic woman. At inception, the UK programme invited all women (age 50 - 64) registered with a GP for screening on a three yearly basis. Over time, the programme has evolved and expanded to now incorporate two-view (Medio-lateral oblique and Cranio-caudal) digital mammography automatically to women aged 50-70, with the option for the over 70 to self-refer. In England, a randomised controlled trial is presently underway to establish if there is a benefit in extending the age range further to 47–73 years (Moser et al. 2011).

Figure 3 Flowchart demonstrating the screening process (adapted from DOH 2013).

Figure 3 provides an explanatory diagrammatic flow chart of the screening process. In the screening programme mammograms are deemed normal (which includes benign findings), or abnormal if there are indeterminate or suspicious findings with

the latter resulting in a recall for further assessment. Clinical recall may also be instigated in women whose mammograms are considered normal but for whom there is documentation at the screening of relevant signs or symptoms of breast cancer.

2.2 NHSBSP Assessment Process/Triple Assessment

The aim of the assessment stage in Figure 3 is to obtain a 'definitive and timely diagnosis' (NHSBSP 2010:1) of potential abnormalities. During the assessment, the workup of cases can involve many options. Ranging from ultrasound only with or without a clinical breast examination (CBE), other cases will require further mammographic views, and ultrasound (U/S) +/- CBE; and in some cases complete triple assessment is undertaken (additional imaging, histological sampling and CBE). This modified triple approach incorporating U/S and Needle Core Biopsy (NCB) (rather than Fine Needle Aspiration Cytology (FNAC)) is evidenced to achieve an accurate diagnosis in 99% of cases (Wai et al. 2013). The individual who is leading the assessment clinic (Consultant Radiologist or Consultant Radiographer as per NHSBSP 2010) makes the decision on which, or if all components of the triple assessment process are required. Subsequently, the decision to discharge a lady back to routine screening also lies with this individual, although it is recommended good practice to obtain a 2nd opinion.

A multi-disciplinary team (MDT) involved in the patient management includes core members consisting of a consultant radiologist and/or consultant radiographer;

breast surgeon; pathologist; oncologist and a clinical nurse specialist in breast care. The MDT meeting offers a forum to discuss individual cases and a collaborative decision is made regarding an appropriate programme of treatment specific to individual patient needs.

2.3 UK Key National Standards

To achieve high quality and consistent outcomes, a number of key standards have been set by the National Coordination Team as a means of monitoring the performance of the NHSBSP. Appendix A lists the current National Minimum Standards for the NHSBSP, which are designed to allow for high detection (sensitivity) of breast cancers while limiting unnecessary recalls and biopsies for benign findings. These quality standards were also introduced to create a sustainable programme with respect to cost, time and workforce implications. Particularly important for this research are selected Standards 2,3,7, 10 and 13 (see Table 1) as they relate to maximising the number of cancers detected (standard 2), and picking up the cancers at an early stage (standard 3). Standard 7 is especially important as the process of arbitration or consensus of discrepant reads is a contributory factor in ensuring that an excess of false positive cases are not recalled. However, this process needs to be performed in a timely manner to ensure that normal results are received within two weeks of attendance for the screening mammogram and a recall to an assessment clinic occurs within three weeks (standard 13). Decision making on discrepant cases is particularly challenging as small cancers are often subtle and display minimal mammographic changes.

Table 1 Selected Current National Minimum Standards for the NHSBSP

Objective	Criteria	Minimum standard	Achievable standard
2. To maximise the number of cancers detected	a) The rate of invasive cancers detected in eligible women invited and screened	Prevalent screen >3.6 per 1,000 Incident screen >4.1 per 1,000	Prevalent screen >5.1 per 1,000 Incident screen >5.7 per 1,000
	b) The rate of cancers detected that are in situ carcinoma	Prevalent screen >0.5 per 1,000	>1.4
	c) Standardised detection ratio (SDR)	Incident screen >0.6 per 1,000 > 1.0	
3. To maximise the number of small invasive cancers detected	The rate of invasive cancers less than 15 mm in diameter detected in eligible women invited and screened	Prevalent screen >2.0 per 1,000 Incident screen > 2.3 per 1,000	Prevalent screen >2.8 per 1,000 Incident screen >3.1 per 1,000
7. To minimise the number of women screened who are referred for further tests	a) The percentage of women who are referred for assessment b) The percentage of women screened who are placed on short term recall	Prevalent screen <10% Incident screen <7% <0.25%	Prevalent screen <7% Incident screen <5%
10. To minimise the number of cancers presenting between screening episodes in the women screened	The rate of cancers presenting in screened women a) in the two years following a normal screening episode b) in the third year following a normal screening episode	Expected standard 1.2 per 1,000 women screened in the first two years 1.4 per 1,000 women screened in the third year	
13. To minimise the interval from the screening mammogram to assessment	The percentage of women who attend an assessment centre within three weeks of attendance for the screening mammogram	>90%	100%

Therefore, to minimise a cancer presenting between the three yearly screening episodes it is imperative that these cases are rigorously evaluated to prevent a cancer case being incorrectly discharged back to routine screening (standard 10).

2.4 Recall Rate

UK guidelines define minimum (<10% Prevalent screen and <7% Incident screen) and achievable standards (<7% prevalent screen and <5% incident screen) for recall rates (objective 7 Table 1). However, there is variance in these standards internationally (Smith-Bindman et al. 2003). European guidelines advise a recall rate of less than 5% for prevalent screens and less than 3% for subsequent screens (Perry et al. 2008). The Dutch Screening Programme is reported to have the lowest recall rates worldwide averaging 1.6%. A low recall rate is not necessarily beneficial because this may be at the expense of a lower sensitivity. Recall rates are also not directly comparable internationally because of differences in the recommended screening interval and age range as demonstrated in Table 2.

Table 2 Demonstrates the international institutional variance in recommended screening age and interval

Institution		**Screening Interval**						***High-risk***	
Society	**Country**	**40-49**	**Interval**	**50-70**	**Interval**	**>70**	**Interval**	**≥40**	**Interval**
American Cancer Society	USA	Offer 40-44 45-54	1 1	Y	1-2	Y	1-2	Y	1
American College of Radiology	USA	Y	1	Y	1	Y	1	Y	1
National Cancer Institute	USA	Y	1-2	Y	1-2	Y	1-2	Y	1
United States Preventative Task Force	USA	Offer or provide the service	1-2	Y	1-2	Y-74	1-2	Y	1
National Breast Cancer Screening Programme	Netherlands	N	-	Y	2	Y-75	2	Y	1
Canadian Task Force on Preventative health care	Canada	N	-	Y 50-69	2	Y 70-74	2-3	Y	1
Agence Nationale d'accreditation et d'Evaluation en Sante	France	N	-	Y	2	Y-74	2	Y	1
National Health Breast Screening Programme	UK	N	-	Y	3	Y	3	Y	1
Swedish National Board of Health and Welfare	Sweden	Y	18-24 months	Y	18-24 months	Y	18-24 months	Y	1

2.5 Mammography Accuracy and Interpretation of Images

It is well documented that mammography is not a perfect screening test (Birdwell 2009) with accuracy being a combination of both sensitivity and specificity. Sensitivity is a measure of the true positives (the proportion of cancer cases correctly

recalled) and specificity a measure of the true negatives (the proportion of normal cases correctly returned to routine screening). A meta-analysis of 9 studies undertaken by Mushlin et al. (1998) estimated a breast screening sensitivity of between 83% and 95%, with a wide range of individual study results (39% to >90%). However, it needs to be acknowledged that this review relates to the pre-digital era. The more recent UK Tommy trial (2015) reported digital mammography sensitivity as 87% but with a specificity of only 58%.

2.6 Reader Performance

A number of confounding factors may also impact on mammographic detection of tumours; some breast cancers are mammographically occult (not visible on mammography), and certain tumour types are notoriously difficult to perceive as they may only exhibit minimal mammographic changes or portray features that overlap with benign and normal variants. A further limitation of conventional 2-Dimensional (2D) mammography is overlapping normal or dense breast tissue, which may obscure underlying lesions (Laming and Warren 2000).

Although there have been technological advances in the equipment (film to digital transition) and techniques, (Computer Aided Detection (CAD), tomosynthesis 3-Dimensional (3D) imaging, and contrast enhanced mammography) the interpretation of the images is still crucially dependent on individual human decision-making skills. Numerous studies have reported the considerable inter-observer variability that exists within mammography reporting, (Berg et al. 2000, Berg et al. 2002, Duijm et

al. 2009 and Skaane et al. 2008) with inconsistent evidence on whether improved reader performance correlates to experience and reading volumes (Theberge et al. 2014, Buist 2011, Barlow 2004 and Miglioretti 2009). False-negative interpretations are a consequence of either perception errors or interpretative errors (Cornford et al. 2005) with fatigue a possible contributory factor (Taylor-Philips et al. 2011). The principal complexity for reporters is balancing the trade-off relationship of attaining a high sensitivity whilst minimising false positives (Wolf et al. 2015), which impact adversely on patient wellbeing (Bond et al. 2015) and represent cost implications in time and resources.

2.7 Reporting of Screening Mammograms

Different reporting strategies are utilised in various regions of the world. In the United States, single radiologist reporting or single radiologist reporting with CAD are commonly employed. Although CAD systems are designed to aid reader perception, this technology remains a topic of continuing research. A recent retrospective review by Lehman et al. (2015:1837) reports that CAD '*does not improve diagnostic accuracy*' and concludes '*insurers pay more for CAD with no established benefit for women'.*

A meta-analysis by Taylor and Potts (2008) concluded that double reading could increase cancer detection rates by 10% compared to single reading at the expenditure of relatively small increases in recall rates. Double reporting by Radiologists specialised in breast screening is the European standard (Perry et al.

2008). Unique to the UK is double reporting undertaken by non-medics (trained radiographers). This was validated in 2012 following an extensive NHSBSP research project (Non-Discordant Radiographer Only Reporting - NDROR) (Bennett et al. 2012). The success of this system has since led researchers in other countries to investigate radiographer interpretation of mammograms (Debono et al. 2015, Torres-Mejia et al. 2015 and Moran and Warren-Forward 2016).

A variation also exists with regards to how double reporting is undertaken. This may be in a blinded /independent manner (the 2nd reader is not aware of the 1st reader's decision) or non-blinded (the 2nd reader is aware of the 1st reader's decision). In a recent review, Klompenhouwer et al. (2016a) has identified that there are only a limited number of studies that have compared the advantages and disadvantages of these reporting strategies, and that arbitration reduces programme sensitivity when blinded double reading is performed. In the UK it is questionable as to whether true blinding occurs in practice; from clinical experience, only one set of assessment paperwork is produced and therefore the 2nd reader is aware when there is discordance and potentially could change their opinion after reviewing the case.

Following the double reading, if there is a concordant negative result (both readers agree that the woman should not be referred) routine screening is instigated in 3-years' time (Figure 4). If there is a concordant positive result (both readers agree that the woman should be referred) recall to assessment is automatically initiated in some centres. Inherent with double reporting is the probability that the two readers

will differ on their opinion regarding recall (Figure 4). Unilateral recall, i.e. women are recalled if either one of the readers recommended further assessment, is associated with high recall rates and therefore discrepant double readings are mainly resolved by some form of arbitration or consensus strategy.

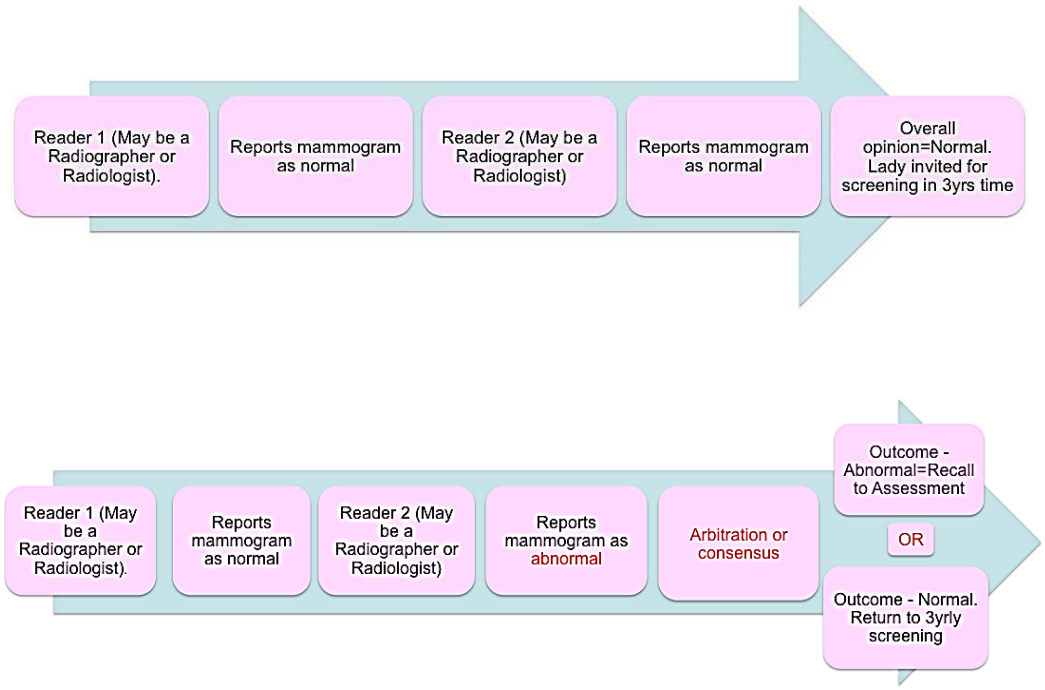

Figure 4 Flow chart demonstrating a normal and an arbitration-reporting scenario in UK practice.

The abnormal read by the second reporter instigates the arbitration or consensus process. The final outcome from the processes is either abnormal and recall to assessment or normal and return to routine screening.

2.8 Resolving Discordant Readings

When there is discordance between the two readers the most common decision methods utilised are arbitration by a third independent reader or some form of consensus review. The independent third reader or lead of the consensus review

currently must be a medic as stipulated by NHSBSP guidance.

Consensus review can involve many differing approaches. Firstly, the initial two readers may attempt to reach an agreement. Secondly, the consensus discussion may include the two original reporters and additional film-reading members within the team. An alternative is consensus review by film readers not including the initial reporters. Finally, complex pathways also exist where both consensus and arbitration are undertaken in the decision-making process.

Kerr and Tindale (2004) and Bankier et al. (2010) describe the complexities of dynamics that exist within consensus discussions where one reader is the dominant and opinions are not equally weighted. The performance-reducing effects of '*group think*' (Bankier et al. 2010: 14) are also an important consideration in consensus where it is evidenced that individuals may change their judgment to what they '*believe others want to hear*' (Bankier et al. 2010: 16).

2.9 Chapter 2 Summary

This chapter has described the reporting processes and standards within breast screening. It has highlighted that there are varying methods for resolving discordant double reporting. Finally, the chapter discussed that while there is national momentum for delegation of arbitration to radiographers, little is currently known about the effectiveness of arbitration versus consensus and whether one strategy

produces improved performance in a breast-screening unit. No systematic reviews to date have been undertaken on arbitration or consensus within mammography reporting and therefore a scoping review is required to establish the nature and volume of evidence to inform the processes and their effectiveness. The next chapter explains the systematic approach utilised to gather this evidence.

Chapter 3. Methodology

The preceding chapters have established the absence of a systematic review of arbitration and consensus processes used in mammography for resolution of discordant reports. Different forms of evidence review can be undertaken, depending on the type of research available. A Cochrane systematic review is considered to be one of the most dependable forms of evidence (Cates, Stovold and Welsh 2014) due to its transparency, objectivity and ability to minimize bias. However, Cochrane reviews are usually tightly focused on quantitative data with confined inclusion parameters. Where the evidence reviewed is qualitative, a narrative review might be selected. This type of review will aim to identify key literature from database searching and may consider grey literature and on-going research but will not extend to '*reference list checking, hand searching journals or contact with experts*' (Booth, Papaioannou and Sutton 2012:83). In the present study it was considered that a scoping review would provide the best overview of the current landscape. A scoping review is described as a method that exceeds a traditional narrative review and can be particularly relevant in generating new insights, gaps within the evidence and can aid to inform future scope and costs of a systematic review (Armstrong et al. 2011). By demonstrating to what extent arbitration and consensus has been investigated, and possibly discovering under-studied areas as well as established sub-fields, recommendations can be made based on a synthesis of evidence from a variety of sources.

This chapter will describe the systematic method used for identifying the literature (searching and screening), and the implications of using varying sources. Data extraction, critique of quality assessment of included studies, and the methodology for synthesis of the study findings are also presented and discussed (Hemingway and Brereton 2009).

3.1 Search Strategy

A broad-based research review (Labin et al. 2012) is a structured method governed by systematic decision rules. Kable et al. (2012) describe a 12-step approach for documenting an effective search strategy, which was considered plausible for this study.

1. **PURPOSE STATEMENT.**

The aim was to review literature pertinent to mammography arbitration or consensus. The primary step was defining the scope of the question, which was challenging and required refinement and development over a period of time. The PICO parameters below were utilised in order to ensure a clearly articulated scope of enquiry (Armstrong et al. 2011). This was considered a significant part of the scoping review process as it influenced the succeeding stages and ultimately the final outcome.

P - Healthcare Professionals:

I – Arbitration/consensus:

C – No arbitration or consensus/highest reader recall:

O – Cancer detection/sensitivity/specificity/PPV:

The PICO framework was not limited to breast screening. Following discussion, it was identified that it would be relevant to assess if there was any evidence relating to arbitration within a symptomatic cancer setting that could be transferable. Particular emphasis was placed on studies evaluating the impact of 3rd reader arbitration or consensus processes incorporating differing reporting strategies (blinded, non-blinded). Various sources of information were searched (databases, conference proceedings, and unpublished data sources) using varying search strategies (combinations of keywords and subject headings) in an effort to exhaust all sources. The aim of the scoping review was to ascertain primary and secondary outcomes of:

Primary Outcome

- What evidence is there to inform arbitration processes within mammography reporting and what evidence is there on their effectiveness?

Secondary outcomes

- To identify potentially useful data sources relating to individual characteristics of the arbitrator in terms of: education and training, experience, frequency and volume of images reported, and decision-making processes utilised.
- To identity gaps in the evidence base and recommend further research if

required

2. SEARCH TERMS

Firstly, keywords and subject headings were collated into broad categories relating to the reporting process, personnel characteristics (e.g. experience, training, volume of films read), decision making and reflective practice, audit and Continuing Professional Development (CPD) (Appendix B). These categories proved challenging as it soon became evident that they were too diverse and extensive. A decision was therefore made to refine the search strategy to make it more realistic and ensure achievable output within the time constraints.

A comprehensive search strategy was formulated by establishing terms originating from the purpose statement (Kable et al. 2012). Concepts of interest (Kable et al. 2012 and Lloyd-Jones and Masterton 2010) were cross-referenced by searching Cochrane reviews for validation (Table 3 below lists the final search terms and variations used). Developing a systematic search strategy initially proved difficult due to variable terminology and indexing across different databases (Derry, Loke and Aronson 2001). Therefore, a combination of keywords, phrases, subject index terms (Thesaurus/MeSH) and the explode function were used. Keyword truncation was also utilised to retrieve results that may have variations in the spelling, plurals and synonyms (Aveyard 2014). Boolean operators 'AND', and 'OR' were applied to focus the most productive search and aid in eliminating inappropriate hits.

To ascertain if the search terms were effective they were tested as recommended by

Kable et al. (2012) to establish if known key papers had been located. Greenhalgh and Peacock (2005) consider that highly structured search strategies across a range of electronic databases can still be unsatisfactory carrying the risk of missing relevant material which may be significant. Betran et al. (2005) report that up to 20 per cent of studies are not retrieved via database searching alone. Multiple studies (Timmins and McCabe 2005 and McGowan and Sampson 2005) advocate the use of an experienced librarian in literature searching. In this study a librarian was utilised to check the search strategies for erroneous spellings or errors in the use of the Boolean operators.

Table 3 Demonstrates the Final Search Terms and Variations Used

Exploded terms	Alternative keywords
Breast neoplasm	breast adj3 (neoplasm* OR carcinoma* OR tumour* OR tumor* OR cancer*.
Mass screening	breast adj3 (scan* OR screen* OR radiograph* OR imaging OR visualise OR visualize OR exam* OR test* OR mammogra* OR routine* OR check* OR diagnos* OR detect*)
Mammography	mammogra* adj3 (scan* OR screen* OR visualise OR visualize OR exam* OR test* OR breast*)
Early detection of cancer	
National Health Service Breast Screening Program	OR "NHSBSP" or "UK breast screen* program*" "NHS breast screen* program*"
Negotiating	arbitration* OR discordan* OR discrepan* OR disparity* OR negotiat* OR disagree* OR conflict* OR differen* OR inconsisten* AND variation* OR consensus* OR uncertain*
Decision making	"decision mak* OR shared decision making" OR "medical decision making" OR "choice behaviour" OR "problem solving" OR "clinical decision analysis" OR "critical think*" OR "decision aids" OR "Task performance and analysis"
Interpersonal communication	

3. SOURCES OF SEARCHES

Databases/Electronic searches: In a scoping review the number of databases searched can be limited and this is recognised as a limitation of this type of review (Levac, Colquhoun and O'Brien 2010). Seven databases were searched in the present study in an attempt to ensure relevant studies were not omitted (Crossan and Apaydin 2010). However, this proved a time-consuming process. The Centre for Reviews and Dissemination (2008) has confirmed comprehensive-scoping reviews can take almost a year to accomplish. The 7 databases were selected based on their relevance to retrieving multi-disciplinary research evidence from the perspectives of medicine, allied health professionals and health care. A list of the databases searched is given in Table 4 below.

Table 4 Databases searched and time frame for searches.

Database	Date range of search	Appendices
PubMed.	No start date restriction - 26th January 2016.	See Appendix C for the full search strategy
MEDLINE	No start date restriction - to 18th January 2016	See Appendix D and E for the full search strategies
EMBASE	1st January 2005 to 18th January 2016).	See Appendix F and G for the full search strategy
CINAHL	No start date restriction - 21st January 2016.	See Appendix H for the full search strategy
Cochrane Library	No start date restriction - 19th January 2016	See Appendix I, J and K for the full search strategy
Scopus	No start date restriction - 26th February 2016.	See Appendix L for the full search strategy
Web of Science	1st January 2005 to 26th February 2016	See Appendix M for the full search strategy

In the above databases, search terms were restricted to title, abstract and keywords within the article or topic subject within Web of Science. A better understanding of

the benefits of specific querying options led to the proximity operator of adjacent 3 being used with search terms in order to retrieve terms within phrases and avoid onerous hits where breast was merely mentioned in the title. Rich Site Summary or Really Simple Syndication (RSS) feed alerts were set up on Scopus and Web of Science for keywords and citations of a current 2015 author; this ensured any updates would be known without the requirement to manually re-check. Within the time limitations it was not possible to re-run all searches. Therefore, an updated search (4th June, 2016) was undertaken in Scopus as this database had retrieved the greatest number of relevant papers. Two further publications were identified that related to DBT and radiographer reporting, but neither were specifically looking at arbitration or consensus processes (Hodgson et al. 2016 and Culpan 2016).

4. INCLUSION/EXCLUSION CRITERIA

Criteria for including or excluding retrieved articles were devised to aid in the retrieval of significant studies and to minimise false positive search results (Bettany-Saltikov 2010 and The Joanna Briggs Institute 2015). Criteria related to the intervention and population characteristics (Bettany-Saltikov 2010) but there was no limitation on study design. These are detailed in Table 5.

Table 5 Inclusion and exclusion criteria

Inclusion criteria
1. Provides an English abstract or summary (to assess content) or the title explicitly demonstrates relevance 2. Specifically mentions breast reporting arbitration, 3rd reader or consensus processes OR 3. Discusses reporting strategies – i.e. single reading, double reading, blinded or non-blinded reading. OR 4. Reports strategies for management of discrepant cases – i.e. higher reader recall, arbitrate all recalls, arbitrate discordant cases only. OR 5. Reports the grade of personnel undertaking the arbitration/consensus/3rd read task i.e. radiologist, radiographer, clinician, surgeon OR 6. Specifically, in relation to arbitration, 3rd reader or consensus mentions any attributes required by the personnel undertaking the task. In particular: *Volumes of films read per annum,* *Number of years' experience of the reporter,* *Attendance at MDT's,* *Decision making skills,* *Audit and reflective practice*
Exclusion criteria
1) Non English-language paper 2) Arbitration, consensus or 3rd reader 'mentioned in passing' but not a significant focus of the article.

5. SEARCH LIMITS

The time frame chosen for the preliminary searches was publications between 1st January 2008 and 26th February 2016. The rationale for the 2008 cut-off for this review was that it would give a 2-year lead in period from when relevant NHSBSP guidance was last revised (2010/2011). The final year 2016 was the year when this review was initiated and therefore represents the current position. An English language restriction was applied to all searches for practical reasons. This was not considered detrimental given the likelihood that high impact papers would be translated into English. Searches were limited to human studies.

Initial searches on EMBASE retrieved small numbers of articles and scanning of the titles revealed that many appeared irrelevant. Therefore, for subsequent searches either the start year was extended to 2005, or no date restriction was applied. This also widened the ability to ascertain if a seminal piece of work was produced prior to the initial date limitation.

Reference lists. A manual search of reference lists in eligible articles was undertaken to look for any other relevant citations; one additional study was identified (Matcham 2004). This was missed due to setting a 2005-year cut-off within two of the databases (see search limits). Original searches by keywords and subject headings were supplemented with a search by key author names of articles published in 2015.

Hand searching relevant journals. A search for key current articles published in the European Radiology and Breast journals was undertaken. Both journals were searched with mammography and arbitration to see if any additional articles could be identified. No further papers were sourced via this route.

Grey literature. A number of studies (Coad, Hardicre and Devitt 2006, Aromataris and Riitano 2014 and Mahood, Eerd and Irvin 2014) have described search methods and sources to locate grey literature. However, Godin et al. (2015: 2) state that no '*gold standard exists for rigorous systematic grey literature search methods'* and their '*transparency and reproducibility'* is frequently inferior to search methods reported from academic database searching. Grey literature was sourced by hand

searching of conference proceedings and doctoral theses in an attempt to avoid positive publication bias (Goldacre 2012). Conference proceedings were selected on the basis that their reporting was of breast imaging (UK or International perspective), and the National Cancer Research Institute (NCRI) was used as a means of the author thinking outside the box to ascertain if modalities other than imaging have transferable arbitration/consensus processes. Searches were limited to 2014 onwards as this allowed a 2-year lead in period for conference abstracts or posters to become a fully published article by 2016. However, since the Society of Breast Imaging/American College of Radiology Breast Imaging Symposium (SBI/ACR) was inaugurated in 2015, 2014 papers/posters could not be retrieved. Therefore, a review of the 2015 programme details only was undertaken. The NCRI cancer conference site proved difficult to search as there was only the option to keyword search and confinement by further filtering was not available. Overall, searching of the grey literature was adaptable and time periods were flexed to available resources. Only one poster presentation from the European Congress of Radiology (ECR) 2014 provided relevant information via this search method.

Table 6 below documents the Grey literature search strategy. English language restriction only was applied.

Table 6 Grey Literature Search

Source	Date range searched	Keywords
OpenGrey	No date restriction- to 26th Jan 2016	"Mammography and decision making" "Mammography and arbitration" "Mammography and reporting"
OpenDOAR	No date restriction- to 26th Jan 2016	Mammography, "Mammography and breast screening"
Ethos	No date restriction- to 26th Jan 2016	"Breast Cancer" "Mammography and arbitration" "Mammography and consensus " "Mammography and reporting" "Mammography and decision making"
Zetoc	Conference Proceedings No date restriction- to 26th Jan 2016	"Mammography and arbitration" "Mammography and consensus" "Mammography and decision making" "Mammography and double reading"
British Society of Breast Radiology (BSBR)	2015 (2014 not available)	
European Congress of Radiology (ECR)	2014 and 2015	
National Cancer Research Institute (NCRI)	2015 2014 (unable to filter search terms)	
Society of Breast Imaging/American College of Radiology Breast Imaging Symposium (SBI/ACR)	2015 (Inaugurated)	

Internet searching. In addition to the academic databases searched, the basic keywords of mammography and arbitration were used to conduct a search on googlescholar.com. As it was not practical to screen all retrieved results from this search, reliance was placed on the Google search engine to list the most relevant results first. It was notable that the Google search retrieved all key 2015 publications that were subsequently only found in Scopus/Web of Science.

Personal contacts. Godin et al. (2015) endorse that professional associates and experts in the field are useful contacts to source unpublished material. In accordance with this guidance key author of publications were contacted from the Netherlands and Germany to ascertain if they were aware of any research in progress or any upcoming publications relating to arbitration/consensus processes in mammography reporting. The UK Royal College of Radiologists (RCR) and the Society and College of Radiographers (SCoR) were also contacted for assistance in locating relevant professional body publications.

To improve the breadth of this element of the review method, the National Lead for Screening Quality Assurance (QA) Services at Public Health England was telephoned to widen the pool of contacts. This led to further contacts being made with (i) the Editor-in-Chief of the *Radiography* journal who checked the editorial system to identify any accepted papers or those currently under review; (ii) the scientific supervisor for the Dutch Reference Centre for Screening; and (iii) the director of the National Retinopathy Screening Programme. Via these contacts the summary from a workshop in America was sourced which was primarily concerned with improving the interpretation of breast images. Greenhalgh and Peacock (2005) describe this adaptable search method as snowball sampling where the strategy evolves with the information and sources retrieved in a receptive manner. These sources along with the British Association for Cytopathology were considered pertinent to the current review as they offered either an international perspective on screening or a comparator against which to benchmark breast-screening reporting processes.

6. DOCUMENTING THE SEARCH AND SELECTION PROCESS

A series of sequential searches was undertaken with search queries and number of hits for each database documented; results were excluded if they retrieved zero results. The complete process is demonstrated in the PRISMA diagram below (Figure 5).

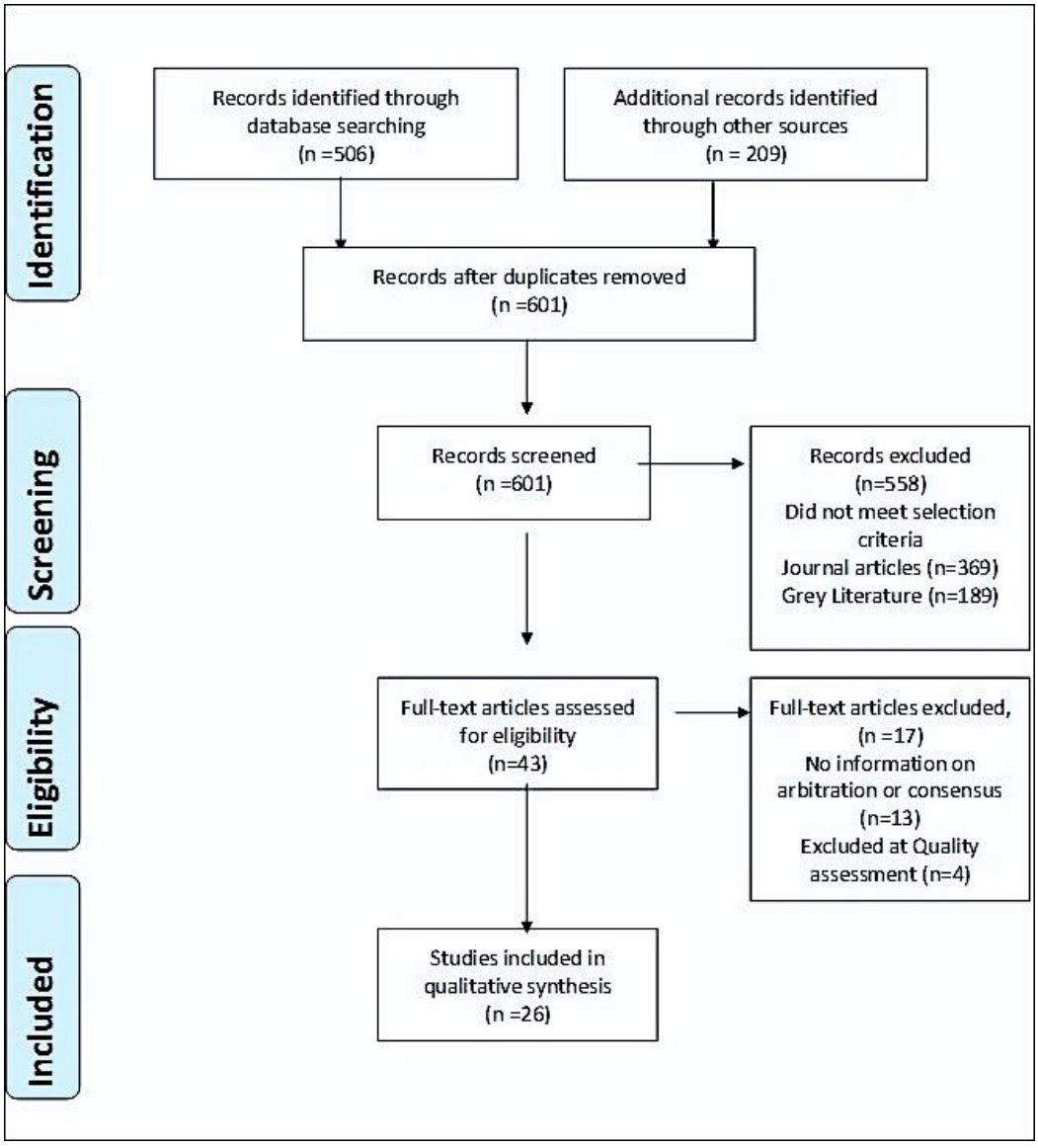

Figure 5 PRISMA 2009 Flow Diagram

An important principle in a scoping review is to document and demonstrate the decisions made in order to show transparency of the search process (Aveyard 2014), which aids in reproducibility and validates the rationale. Studies were saved in an electronic folder and the online reference manager (EndNote) was utilised to remove any duplicates.

7. TEST RELEVANCE OF RETRIEVED ARTICLES

A three-stage process was utilised for filtering the array of published papers and grey literature retrieved (Bettany-Saltikov 2010).

First stage: selection of literature for entry into review - First stage selection was undertaken based on an analysis of the titles and/or abstracts or summaries of all material identified through the various search strategies (Aveyard 2014). This resulted in exclusion of any papers/posters that were clearly not relevant and in the selection of potentially eligible articles, as well as determining the focus of the available literature.

Second Stage: article/grey literature selection - In the second screening stage, two reviewers independently screened abstracts for all retained literature, against the agreed inclusion and exclusion criteria. Any disagreement was resolved after retrieval and review of the full text (five articles identified and arbitrated).

Third Stage: Inclusion and exclusion criteria applied to full article - The full text of all potentially eligible peer-reviewed papers /grey literature items were retrieved.

Two reviewers independently examined these to determine whether the article provided sufficiently detailed evidence against the inclusion criteria, confirming that the abstract was not misleading. A third reviewer resolved any disagreements over the eligibility of a particular study (no articles identified).

8. SUMMARY TABLE OF INCLUDED ARTICLES

Articles that met the inclusion criteria were documented in a customised data extraction form (Table 7) (Maslin-Prothero and Bennion 2010 and Cummings et al. 2010). This was designed to capture and summarise key information and major findings (The Joanna Briggs Institute 2015). The data extraction form enabled raw data from multiple disparate studies to be amalgamated and compared, aiding in pattern recognition and providing a '*rapid and succinct summary of the literature for review*' (Kable et al. 2012: 878). A decision was made to undertake data extraction separate to quality assessment in order to ensure that an overview of study details was available prior to concluding the overall judgement on quality.

Table 7 Articles included in the review

1. Klompenhouwer et al (a) (2015) ***Netherlands – Quality CASP criteria met***						
Research question/aim	**Study design**	**Reporting/ Arbitration strategies**	**Method Data Collection Sample size**	**Data Analysis/ Metrics**	**Main findings/results**	**Strengths/ Weaknesses**
Effect of arbitration by a 3rd reader discrepant reading for blinded and non-blinded double read-screening mammography Mammographic abnormalities and tumour characteristics of cancers missed after arbitration.	Retrospective review Quantitative design	Double reading Blinded and non-blinded alternated on a monthly basis Discrepant readings were always recalled Retrospectively reviewed by a 3rd radiologist – blinded to outcome. Used BI-RADS classification	Consecutive series of 84,927 mammograms 1st July 2009 -1st July 2011. 3 units – 12 radiologists, 1-15 years of screening mammography experience. FFDM Discrepant cases randomly assigned	Recall rate, cancer detection rate, proportion of BI-RADS 0 among all recalls, PPV, programme sensitivity. Cancers not recalled after arbitration by a third reader calculated as interval cancers. Independent-sample t- test. (95 % CI). Chi square and Fisher's exact tests - differences in tumour and mammographic characteristics of the reading strategies, differences in surgical treatment. P-value < 0.05	Discrepant readings =57.2 % blinded vs. 29.1% non-blinded, (p< 0.001), **Blinded double reading, arbitration**= 1. Decreased recall rate (3.4 to 2.2 %, p< 0.001) 2**.decreased sensitivity (83.2 to 76.0 %, p = 0.013)** 3.**No influence** on cancer detection rate (CDR; 7.5 to 6.8 per 1,000 screens, p = 0.258) 4. Increased the PPV; 22.3 to 31.2 %, p <0.001). **Non-blinded double reading, arbitration =** 1. Decreased recall rate (2.8 to 2.3 %, p < 0.001) 2.increased PPV (23.2 to 27.5 %, p=0.021) 3.no affect on affected CDR (6.6 to 6.3 per 1,000 screens, p=0.604) 4.no affect on sensitivity (76.0 to 72.7 %, p=0.308). No differences in the proportion of DCIS, smaller tumours, lymph node Involvement or advanced tumours among SDCs and cancers missed at arbitration. Invasive cancers with axillary lymph node	Weakness – Acknowledged by the author arbitration outcome did not affect "real-life". Discrepant cases were recalled regardless. Therefore, the arbitrator's role did not have clinical implications for the screening. Strengths - Waited 2 yr. screening period to capture "interval cancers". True sensitivity calculated. Prior films available Number of radiologists with variable experience reflects clinical practice Large case series

					metastasis were less often seen among cancers Missed at arbitration (20.3 % vs. 11.1 %, p<0.001	

2. Klompenhouwer et al (b) (2015) *Netherlands* - Quality CASP criteria met						
Research question/aim	**Study design**	**Reporting/ Arbitration strategies**	**Method Data Collection Sample size**	**Data Analysis/ Metrics**	**Main findings/results**	**Strengths/ Weaknesses**
Evaluate PPV, discrepant rate, and characteristics of BI-RADS 0 recalls screening program. Determine the effect of arbitration by a 3rd reader of discrepant BI-RADS 0 readings at both reading strategies.	Retrospective review Quantitative design	Double reading Blinded and non-blinded -alternated on a monthly basis. Discrepant readings were always recalled Retrospectively reviewed by a 3rd radiologist – blinded to outcome. Used BI-RADS classification	Consecutive series of 84,927 1st July 2009 – 1st July 1 2011. 3 units – 12 radiologists 1-15 years of screening mammography experience. FFDM Discrepant cases randomly assigned	Chi square or Fisher exact test - differences in categorical variables PPV of recall of BI-RADS categories. Cancers not recalled after arbitration by a third reader were calculated as interval cancers. Continuous variables - double sided t-test for independent samples P-value < 0.05	Arbitration of discrepant BI-RADS 0 recalls = lowered recall rate (from 3.4% to 2.8% at blinded double reading, p < 0.001, and from 2.8% to 2.5% at non-blinded double reading, p 1⁄4 0.008), without a decrease in cancer detection rate (from 7.5‰ to 7.3‰, p 1⁄4 0.751, and from 6.6‰ to 6.5‰, p 1⁄4 0.832, respectively) and program sensitivity (from 83.2% to 81.2%, p 1⁄4 0.453, and from 76.0% to 74.6%, p 1⁄4 0.667, respectively). Arbitration would have significantly increased the PPV at blinded double reading (from 22.3% to 26.3%, p 1⁄4 0.015). 13 cancers missed by arbitration - overall decrease in cancer detection rate is very small, 0.1-0.2% at both reading strategies No differences in mammographic and tumour characteristics of BI-RADS 0 Recall at blinded and non-blinded reading	Weakness – Acknowledged by the author arbitration outcome did not affect "real-life". Discrepant cases were recalled regardless. Therefore, the arbitrator's role did not have clinical implications for the screening. No cost-effectiveness Strengths– waited 2 yr. screening interval to capture "interval cancers". Large case series Number of radiologists with variable experience reflects clinical practice

3. Hofvind et al (2009) *Norway* - Quality CASP criteria met						
Research question/aim	**Study design**	**Reporting/ Arbitration strategies**	**Method Data Collection Sample size**	**Data Analysis/ Metrics**	**Main findings/results**	**Strengths/ Weaknesses**
Analyse discordant and concordant screen detected breast cancers using independent double reading with consensus**.** Arbitration only if consensus not reached by initial reporters	Retrospective review Quantitative design	Double reading Blinded reading Score 1-5 1, normal; 2, probably benign; 3, indeterminate; 4, probably malignant; and 5, malignant. Initial score of 2 or higher by either reader = a consensus meeting Initial score of 3 or higher – can't be dismissed without agreement from initial reporter Arbitration only if consensus not reached by initial reporters	1 033 870 prevalent and incident screens 5611 screen detected cancers (DCIS + invasive) 1996–2005 Radiologists Average experience = 4.3 years (range, 1–11 years), average volume for the whole study period (9 yrs.) = 19, 745 screening mammograms range, 525–107 161. SFM= 97% FFDM = 3%	Differences in rates and proportions tested with a x2 test. All tests were two-sided. *P* values <0.05. Logistic regression to estimate the odds that a discordant cancer was associated with mammographic density. Odds ratios (ORs) with 95% CI - adjustment for age at screening and prevalent vs. incident screening K Statistics - for agreement between two readers. Unweighted K values for 2 x 2 table analyses (positive and negative scores) Quadratic weighting for five-point interpretation scale. Observer agreement, k values < 0.20 =poor agreement; 0.21– 0.40, fair agreement; 0.41– 0.60, moderate agreement; 0.61–0.80, good agreement; and more than 0.81, very good agreement SPSS	Discordant scores = 5.3% Concordant positive scores = 2.1% At consensus, 66.8% (36 380 of 54 447) of the discordant and 17.9% (3932 of 21 928) of the concordant screenings were dismissed. Recall rate = 3.5% 23.6% (1326 of 5611) of CA had discordant interpretation**.** Varied from 16.9% (148 of 874 cancers) to 28.6% (265 of 928 cancers) according to county 117 interval breast cancers were diagnosed among the 40 312 screenings that were dismissed at consensus = 6.5% of all interval cancers.	Weakness – Acknowledged by author - Don't know if score correlates with actual CA and if the 2 reporters recalled for the same abnormality as quadrant and lesion characteristics not specified at initial interpretation 2 radiologists read less than 500 screening mammograms during 1 year in study period. Against the exclusion criteria No cost effectiveness Strengths - Large case series Specialist and general radiologists – representative of a community setting, but no information provided on the amount of time non-specialists dedicate to breast

4. James and Cornford (2009) *UK*. - Quality CASP criteria met						
Research question/aim	**Study design**	**Reporting/ Arbitration strategies**	**Method Data Collection Sample size**	**Data Analysis/ Metrics**	**Main findings/results**	**Strengths/ Weaknesses**
Can computer-aided detection (CAD) act as an arbitrator of discordant double-reading opinions, replacing the need for an independent 3rd film reader.	Retrospective review Quantitative design	Double reading Not completely blind Original arbitration by independent 3rd reader – radiologist Arbitration Mammograms digitised and analysed by CAD system – compared to radiologist CAD algorithms set to operate at a detection sensitivity of 88% for masses and 95% for micro calcifications.	240 cases underwent arbitration from 16,629 cases July 2003-April 2004. 5 radiologists, 1 research fellow, 1 radiographic film reader Radiologists experience ranged - 5-18 yrs. radiographer - 5 years	Statistical significance - McNemar test to take into account the matched nature of the data.	Arbitration cases accounted for 22% (112/518) of total cases recalled for assessment. 47% cases recalled to assessment following the opinion of the arbitrator 21 cancers in arbitration set, 13 diagnosed at the time of the original screening mammogram, 8 diagnosed subsequently. 3 were not the arbitrated lesion, **5 were** – 2 of these were assessed and returned to RR. CAD correctly prompted in these 5 cases. 2 cancers recalled by arbitrator and not CAD Independent 3rd reader recalled 15/18 (83%) of the cancers that corresponded with the arbitrated lesion. CAD as the arbitrator would have recalled 16/18 (89%) of the cancers that corresponded to the arbitrated lesion. CAD= **significant increase in normal women being recalled to assessment in the arbitration group (P < 0.001). <u>Extra 50</u>** recalls. Recall rate increase from 3.1 to 3.4%; increase of 10%. Overall –No. Of cancers detected were broadly similar with 1 additional cancer recalled by CAD	Strengths - Reporters included radiographer – represents current UK practice Weakness – acknowledged by author -Small number of cancers in the series (18) Retrospective - can only give an indication as to the potential effect of CAD acting as an Arbiter No cost effectiveness Not completely blinded reading - may influence the proportion of discordant cancers.

5. Mucci et al (1999) *UK* - Quality CASP criteria met						
Research question/aim	**Study design**	**Reporting/ Arbitration strategies**	**Method Data Collection Sample size**	**Data Analysis/ Metrics**	**Main findings/results**	**Strengths/ Weaknesses**
Experience of double reading – breast screening 3rd person arbitrator	Prospective study Quantitative Design	Double reading Non-blinded 3rd reader decision= final decision. Non-blinded.	398 arbitration cases 1992-1994 3 radiologists	% Calculated for recall rates	398 arbitration cases - final reader recalled 196 (49%) and returned 202 (51%) to routine recall – 1 true interval CA subsequently Of 196 assessed - 4 malignant. Estimated cost saving by arbitration £20,000– 202 women returned to normal screening Assessment episode is £101, 3rd read=£1 (1999 figures) 3rd reader =reduction in no. Of recalls and no reduction in cancer detection.	Weakness – acknowledged by author -non-blinded 2nd reader knew the opinion of the first and was influenced. Therefore, underestimate the benefits of double reading to cancer detection. Strength - 3rd reader was aware of the opinion of the first two; simply asked to arbitrate on the action to be taken on an identified lesion – real clinical practice
6. Liston and Dall. (2003) *UK* - N/A for CASP -audit						
Method for assessing performance of new readers Arbitration	7yr Audit	Double read Non blinded Independent review by 3rd reader. Majority opinion is acted upon.	1/4/95 - 31/3/02 5 radiologists Varying experience	% Calculated for Cancers incorrectly returned to RR by 1st and 2nd reader Total no. Of cancers detected through double reading	The % of cancers detected with double reading + 3rd reader arbitration varied each year -3.6 and 11.4% Overall 87 (8.1%) of the 1072 cancers were detected following 3rd reader arbitration.	Strength - Robust audit

7. Cornford et al (2005) *UK* - Quality CASP criteria met						
Research question/aim	**Study design**	**Reporting/ Arbitration strategies**	**Method Data Collection Sample size**	**Data Analysis/ Metrics**	**Main findings/results**	**Strengths/ Weaknesses**
Compare the mammographic background pattern, mammographic and pathological features of screen-detected cancers diagnosed following arbitration of discordant double reading opinions with cancers diagnosed following concordant double reading.	Retrospective review Quantitative design	Double reading Not entirely blinded 3rd reader arbitrator – had final decision. Independent decision – but not blinded to initial reports	April 2002 - December 2003 32,613 screened 431 arbitration cases 5 radiologists, 1 research fellow 1 radiographic film reader. Radiologists' experience ranged from 5–18 yrs. Film reader =5 yrs. experience.	Chi-square and Fisher's exact tests. Comparison of normally distributed, continuous variables, such as patient age, was analysed with unpaired t-test with Stat- View	287 malignancies. 38 (14%) had undergone arbitration and 249 (86%) had concordant double reading. 50% of arbitrated cases were recalled for assessment -38 malignant [PPV=18%]. Arbitration cases accounted for 20% of the total recalls. Arbitration group – 1st reader did not recall 27 malignancies; 2nd reader did not recall 11 malignancies. Arbitration group =27 invasive cancers and 11 DCIS. Concordant group = 196 invasive cancers and 47 DCIS. = No significant difference between 2 groups. No significant difference in proportion detected through a first or subsequent screen in the two groups (p<0.7). Cancers detected following arbitration were more likely to manifest as parenchymal	Weakness - 2nd reader not entirely blinded – may affect cancer detection rates, but does reflect normal clinical practice. Only 2/5 radiologists as arbitrators Only 1-year f/u – too short to assess all interval cancers Strength - Arbitrator not blinded –reflects normal clinical practice Reader workforce representative of UK practice, radiographer included. All with substantial experience.

					distortions p<0.001 and less likely to manifest as spiculate masses p<0.014). Less likely to be detected in fatty breasts p<0.01). Were smaller (p<0.045). Lobular cancers were commoner in the arbitration group, although this was of borderline significance, p<0.057 Estimated -11% more cancers are detected as a result of double reading with arbitration compared with single reading alone, after taking into consideration second reader bias.	

8. Caumo et al (2011) *Italy* - Quality CASP criteria not met						
Research question/aim	**Study design**	**Reporting/ Arbitration strategies**	**Method Data Collection Sample size**	**Data Analysis/ Metrics**	**Main findings/results**	**Strengths/ Weaknesses**
Role of arbitration by 3rd reader of discordant double readings to reduce recall rates	Prospective Quantitative Design	Double reading 3rd reader – **only 1 person used** Assessment performed irrespective of arbitration results	15/9/09 - 15/1/10 298 arbitrated cases **Only 1 radiologist as the arbitrator** >30 years' experience FFDM	Observed differences were checked by the chi-square ($\chi 2$) test, p *value <0.05.*	Recalls rate at double reading =6.8%. 230 (43.5%) were concordant + 298 (56.5%) were discordant. After arbitration classified – 216 (72.4%) negative + 82 (27.6%) positive 43 (18.6%) cancers were in concordant group 6 (2%) discordant recalls 5 were recalled 1 CA would have not been recalled Arbitration = reduced 216 assessment procedures (2.8% absolute, 40.9% relative reduction of recall rate) missed 1 CA (0.13‰ absolute, 2.0% relative reduction of cancer detection rate). Arbitration had a sensitivity of 83.3% Arbitration cost calculated as adding 3rd reader = 0.25 euros Assessment cost = 67.4–110.4 euros per Discordant readings, often resolved by additional views or ultrasound = lower cost to concordant recalls, more likely to require a biopsy. Based on above - Arbitration cost = 74 euros, 216 spared assessment =14,558.4–23,346 euros. Bias adjusted for by doubling the cost per mammography reading to 0.50 euros and by reducing the cost per assessment procedure to 50 euros. Arbitration = saved cost of 10,651 euros.	Weakness - Only 4-month period in study Only used 1 radiologist as the 3rd reader who had extensive experience >30yrs –not representative of the majority All cases were assessed and therefore the arbitrator's role did not have clinical implications for the screening. Comment Author acknowledged, "some imprecision of cost estimates might have occurred". 1st reading-cost estimates calculated from an excellence centre – does not reflect the average National scenario.

9. Ciatto et al (2005) *Italy* - Quality CASP criteria not met						
Research question/aim	**Study design**	**Reporting/ Arbitration strategies**	**Method Data Collection Sample size**	**Data Analysis/ Metrics**	**Main findings/results**	**Strengths/ Weaknesses**
Effectiveness of arbitration of discordant double readings in mammography screening	Prospective study Quantitative design	Double reading Does not state if blinded Arbitration – 3rd reader	2000–4, 1217 cases 9 radiologist readers 7 radiologist arbitrators Experience - mammograms (at least 10,000 mammograms read and at least three years of screening experience).	% Of sensitivity/ NPV /recall rates	1217 discordant double readings 476 cases (39.2%) arbitrated to assessment, detecting 30 cancers (6.3%). Of 741 negative arbitrations (60.8%), 311 F/U thus far = 2 cancers (0.64%) occurred in the site previously suspected at one of the two independent readings. Assumed Arbitration sensitivity = 86.3% NPV 99.3%. Arbitration reduced the overall referral rates from 3.82% to 2.59% (relative decrease 32.1%). false-negative arbitration, cancers detected per 1000 women screened would decrease from 4.58 to 4.50 (relative decrease 1.7%). 2005 standards: cost per arbitration = 4 euros, assessment 147 euros. For every 1 cancer missed due to arbitration - 151 recalls and 21,248 euros would have been saved, whereas the saved cost per screened woman due to arbitration was 1.72 euros.	Weakness - Only followed up 42% so far so estimated cancer detection rate. Rates transposed to full population screening to give the sensitivity/NPV recall etc. NOT continuous cases -limited to periods when radiologists were available to perform a 3rd third read Strengths - Acknowledged by author - cost analysis cannot be generalized to any other setting, as costs may vary substantially from one country to another and possibly among different centres.

10. Cawson et al (2009) *Australia* - Quality CASP criteria met						
Research question/aim	**Study design**	**Reporting/ Arbitration strategies**	**Method Data Collection Sample size**	**Data Analysis/ Metrics**	**Main findings/results**	**Strengths/ Weaknesses**
Compare double reading and arbitration (BP) for discordance, with CAD Invasive CA only	Retrospective cases. Quantitative design	1. Single read 2. CAD-assisted single Reading 3. Double reading - blinded	January 1998 to December 2001 Total 1569 cases 157 randomly selected double-read Invasive cancers were mixed 1:9 with normal cancers. 2 Radiologists Reader A - (**>**5000 cases/year) 7 years screening experience Reader B - senior radiology Trainee - 6 months training 3rd reader (10 years' experience Reading **>**5000 cases/year) Verified whether lesions recalled by the readers corresponded to cancers.	95% CI Comparison of sensitivities of 2 reading methods - Stata 'prtest' T-tests - to compare mammographic diameters. ROC curves plot sensitivity against specificity	The CAD system was highly Sensitive (93%, 95% CI 87.8–96.5), detecting many cancers overlooked by the readers, but the readers rejected most TP prompts CAD prompts are numerous and mostly FP. BP sensitivity = 90.4% CAD+RA sensitivity =86.6% (P = 0.12) CAD+RB 94.3% (P = 0.14). CAD-RB specificity was less than BP (P = 0.01). After CAD, reader's sensitivity increased 1.9% and specificity dropped 0.2% and 0.8%. **Arbitration decreased specificity 4.7%.** ROC analysis = BP accuracy better than CAD+RA, borderline significance (P = 0.07), but not CAD-RB. Cancers recalled after arbitration (P = 0.01) and CAD-R (P = 0.10) was smaller.	Weakness - Prior mammograms were not available – may affect a reader's decision to recall Relatively high ratio of cancers to normal cases in the test set Readers had no prior inexperience with CAD Don't know what level of sensitivity the CAD system was set to. Only 2 readers utilised. Trainee as 1 of readers although sensitivity higher than experienced radiologist Strengths - Excluded cancer cases that were previously detected by the readers to eliminate bias due to recollection. Waited 2 yr. screening interval to capture "interval

					No difference in cancer size or sensitivity between reading methods was found with increasing breast density. CAD-R and BP sensitivity and cancer detection size were not significantly different.	cancers".

11. Taylor and Potts (2008) *UK*.- Quality CASP criteria met						
Research question/aim	**Study design**	**Reporting/ Arbitration strategies**	**Method Data Collection Sample size**	**Data Analysis/ Metrics**	**Main findings/results**	**Strengths/ Weaknesses**
Compare single reading with CAD to single reading without CAD Compare double reading to single reading Arbitration and consensus.	Systematic review	1.Single reading 2. Double reading 3. Consensus 4. Arbitration studies	1991-2008 27 studies in total	Meta-analysis using the 'metan' command in Stata 8.2. Becker–Balagtas marginal estimated odds ratios Fixed effects models (using the Mantel–Haenszel method), random effects models (DerSimonian and Laird method) when heterogeneity as high.	Heterogeneity within each of the groups for recall rates. Arbitration/consensus studies, $p < 0.001$ Overall, arbitration studies show a **decrease in recall rates,** but two, including one of the largest studies, show a significant **increase.** Double reading – recall rates with arbitration - overall pooled estimate for the odds ratio is 0.94 (95% CI: 0.92, 0.96; **v**2 (1) = 30.1, $p < 0.001$). As a risk difference, this is a reduction of 2.67 per 1000 (95% CI: –1.72, –3.62; $z = 5.49$, $p < 0.001$). Random effects models - pooled estimate for arbitration/consensus studies is lower, but a larger confidence interval means that the result is marginally not significant (OR = 0.87; 95% CI: 0.75, 1.02; $z = 1.67$, $p = 0.095$). Double reading with arbitration **increased detection rate**	Strengths - Met all the CASP criteria – transparent methodology

					(confidence interval (CI): 1.02, 1.15) and decreases recall rate (CI: 0.92, 0.96). Double read – cancer detection rates with arbitration/consensus – overall pooled estimate for the odds ratio is 1.08 (95% CI: 1.02, 1.15; **V** 2(1) = 6.2, p = 0.012) and the risk difference is 0.44 per 1000 (95% CI: 0.10, 0.79; z = 2.50, p = 0.012). For double reading with arbitration, the number needed to treat is 2222 women screened for each additional cancer detected. **CAD** does not have a significant effect on cancer detection rate (CI: 0.96, 1.13) and **increases recall rate** (95% CI: 1.09, 1.12). Evidence that double reading with arbitration enhances screening is stronger than that for single reading with CAD.	

12.Groenewoud et al - (2007) *Netherlands* - Quality CASP criteria not met						
Research question/aim	**Study design**	**Reporting/ Arbitration strategies**	**Method Data Collection Sample size**	**Data Analysis/ Metrics**	**Main findings/results**	**Strengths/ Weaknesses**
Compare reporting strategies – cost effectiveness 1.decision by one of the readers 2. Refer if both agree (consensus) 3.arbitration by a 3rd reader	Retrospective cases Quantitative design	Blinded reading 1.single reading; 2.double reading with referral if any Reader suggests 3. Double reading with referral only if both radiologists agreed	26 radiologists volunteered 10 read all films 18 read sub-sets Test set of 500 cases 250 controls 125 screen-detected Cancers 125 interval cancers	MIcrosimulation SCreening ANalysis (MISCAN) to estimate cost-effectiveness	Double reading with referral if any reader suggests resulted in a 1.03 times **higher sensitivity** (76.6%) and a 1.31 times **higher referral rate** (1.26%) than double reading with consensus. **Figured assumed – extrapolated** Assuming a relative increase of the detection rate by 2% and a relative increase of the referral rate by 30% double reading with referral if any reader suggests is comparably cost-effective to double reading with consensus (**e** 2,168 and **e** 2,207 per life-year gained, respectively). Control cases concordant =90.2% 89.4% both readers=normal case. 0.8% they both recommended referral. Cases concordant =75.2% 59.3% both readers=normal case 15.9% they both recommended referral. Of all readings by the 153 radiologist pairs, 17.7% were discrepant. Referral rates were highest with decision-making by consensus =73.8% decision by 1 reader = 57.4% arbitration = 52.7%	Weakness - Experimental setting not reflective of daily practice Used published regional Data to estimate the distribution of concordant and discrepant readings Assumed that each referral of a case would lead to the diagnosis of cancer

13. Lång et al (2016) *Sweden* - Quality CASP criteria met						
Research question/aim	**Study design**	**Reporting/ Arbitration strategies**	**Method Data Collection Sample size**	**Data Analysis/ Metrics**	**Main findings/results**	**Strengths/ Weaknesses**
Performance of one-view digital breast tomosynthesis (DBT) in breast cancer screening. Arbitration	Prospective one-arm single-institution study Quantitative design	Blinded reading Double reading and scoring Arbitration = **at least two readers** decided on recall irrespective of the score on the other modality Conventional 2 view DM 1 view (MLO) DBT	January 2010 to December 2012 Aim for 15,000 this study reports first half - 7500 cases 6 radiologists 5 = > 10 years' experience 1 reader =< 10 years' experience Mean 26 years, range 8 to 41 years) Individual training in interpretation of DBT images	McNemar's test for paired data of DBT and DM screens for differences in detection and recall rates with 95 % CIs. Differences in characteristics between cancers detected solely by DBT and all DM-detected cancers tested using chi-2 test and Fisher's Exact test, if the sample size was small. Analyses -Stata software (version 13). 80% power ROC analysis	Recall rate after arbitration was 3.8 % (3.3 to 4.2) for DBT and 2.6 % (2.3 to 3.0) for DM ($p<0.0001$). The PPV was 24 % for both DBT and DM.	Strength - Large prospective cohort Readers had DBT experience Weakness - Interim analysis - does not have 80% power at this stage

14.Duijm et al, (2004) *Netherlands* - Quality CASP criteria met						
Research question/aim	**Study design**	**Reporting/ Arbitration strategies**	**Method Data Collection Sample size**	**Data Analysis/ Metrics**	**Main findings/results**	**Strengths/ Weaknesses**
Determine the value of arbitration by a panel for discordant screen reads Arbitration and consensus	Prospective design Quantitative design	Blinded reading Double reading Mutual consensus between 2 readers. Persistent discordance went to arbitration panel = 3 Radiologists different to original reporters Referred to assessment if at least one arbitration member considered necessary. 3 panel radiologists aware of discordant reads but Blinded to results of the other arbitration panellists.	July 1, 1998, and January 1, 2001. 65,779 cases screened 332 discrepant cases 8 radiologists Experience in reading screening mammograms varied from 15 to 36 months (mean, 31 months).	% Or recall rates, cancer detection rates	Concordant referral = 498 (0.8%) of 65,779 screened Concordant normal = 64,949 (98.7%) women. Initial Discordant = 332 (0.5%) cases. After a mutual consultation, disagreement persisted 183 (0.3%) mammograms. Arbitration panel referred 89 of 183 cases. CA = 20 (22%) cases. 3 (3%) of the 94 not referred by the panel, breast cancer was detected at the site of previously discrepant mammographic findings seen at subsequent screening performed 2 years later. **Arbitration panel missed** If all 183 discrepant cases had been referred, the referral rate would have increased from 0.8% to 0.9% at subsequent (incident) screenings and from 1.5% to 1.7% at initial screenings. At subsequent screenings, the number of cancers detected per 1,000 women screened would have increased from 4.4 to 4.5.	Strength - 2yr. screening interval complete Able to assess no. Of interval cancers. Prior films available Blinding of arbitrator to other arbitrators

***15*. Khoo et al *(2005) UK* - Quality CASP criteria met**						
Research question/aim	**Study design**	**Reporting/ Arbitration strategies**	**Method Data Collection Sample size**	**Data Analysis/ Metrics**	**Main findings/results**	**Strengths/ Weaknesses**
Recall and cancer detection rates with and without computer-aided detection (CAD) of discrepant cases-screening Consensus	Prospective design Quantitative design	Blinded reading Double reading - by at least 1 radiologist Each reader viewed current and available prior mammograms for each case – recorded an opinion CAD prompts for the current mammograms displayed - reader reassessed the prompted areas before recording a revised assessment Arbitration cases - discussed by an additional 2 consultant radiologists reviewed current/prior images, CAD prompts, and proforma	March 21, 2003, and January 9, 2004, 6111 case – images digitized 1639 cases arbitrated 12 readers – 7 radiologist + 5 radiographers 4 to 23 years' experience - Mean of 11 years	Relative sensitivity was calculated for each of three protocols (i.e., single reading, single reading with CAD, and double reading) Recall and cancer detection rates 95% CI Estimates for the time spent on arbitration per reader by monitoring time taken and number of cases arbitrated over a 3-week period	62 CA detected. CAD prompted 51 (84%) of 61 radiographically detected cancers. Of 12 cancers missed on single reading, 9 were correctly prompted; 7 prompts were overruled by the reader. Sensitivity Single reading was 90.2% Single reading with CAD was 91.5% Double reading without CAD was 98.4% **1639 cases arbitrated** **39% recalled to assessment** **61% - routine recall** **More women were allocated to arbitration when mammograms were read with CAD -13.8% to 10.5% non CAD** More women were recalled for assessment in the CAD group -6.1% to 5% non-CAD Cancer detection rates = no difference	Strength - Prior mammograms available if possible Weakness - The sensitivity the CAD system was set to is not mentioned True false-negative rate – can't be calculated 3 years of follow-up needed. Unable to assess if any cancers were arbitrated to normal and have developed since

16. Posso et al (2016) Spain - Quality CASP criteria met						
Research question/aim	**Study design**	**Reporting/ Arbitration strategies**	**Method Data Collection Sample size**	**Data Analysis/ Metrics**	**Main findings/results**	**Strengths/ Weaknesses**
Costs and health-related outcomes of double versus single reading of digital mammograms in a breast cancer-screening programme. Arbitration and consensus	Retrospective cases	Blinded Double reading Discrepant reads first discussed by consensus persistent discrepant cases went for arbitration by 3rd third senior radiologist	June 2009 until May 2013, 57,157 cases 4 radiologists (2010 value for costings)	Student's t-test, Chi-square Test, and Fisher exact test. Statistical tests were two sided P values < 0.05 Analyses were performed using Microsoft Excel (2011) and IBM SPSS software version 21.0 (SPSS, 2013).	Discordance between radiologists in 4.5 %(**N**= 2,556) cases 98.1 % (**N**= 2,508) resolved by consensus and 1.9 % (**N** = 48) by arbitration **Estimate affect** Cost. Double reading without consensus and arbitration was 14 % (€ 36,341) more expensive than double reading with consensus and arbitration. Health-related outcomes. Double reading without consensus and arbitration had 1.5 % more false positive results than double reading with consensus and arbitration ($p < 0.001$). Both reading strategies had similar cancer detection rates ($p = 0.986$). Double reading with consensus and arbitration was 15%(Euro 334,341) more expensive than single reading with first reader only. False-positive results were more	Weakness - No interval cancer rates -results are not conclusive Did not calculate the cost-effectiveness of reading strategies

					frequent at double reading with consensus and arbitration than at single reading with first reader only (4.5 % and 4.2 %, respectively; P <0.001). Single reading could reduce the frequency of false positive results without changing the cancer detection rate.	

17.Dinnes et al (2001) UK - Quality CASP criteria met						
Research question/aim	**Study design**	**Reporting/ Arbitration strategies**	**Method Data Collection Sample size**	**Data Analysis/ Metrics**	**Main findings/results**	**Strengths/ Weaknesses**
Compare double reading with single reading of mammograms for screening accuracy, patient outcomes and costs. Arbitration and consensus	Systematic review	Single reading and Double reading For double reading recall policies 1. Recall if 1 suggests 2. Arbitration 3.consensus 4. Mixed Mixture of blinded and non-blinded Double reading	April 1991 -July 1999 10 cohort studies met inclusion criteria Only 3 studies evaluated for sensitivity and specificity		Consensus or arbitration or a mix of the two, decreased recall rates (by between 61 and 269 per 10,000 women screened). Insufficient evidence was available to detect any pattern in cancer detection according to recall policy. Specificity increased with consensus or mixed recall. Unable to analyse cost effectiveness as significant variation between the organisation of services from different countries Unable to quantify a difference on cancer detection rates from the results.	Strength - Met CASP criteria

18. Skaane et al (2013) *Norway*- Quality CASP criteria met						
Research question/aim	**Study design**	**Reporting/ Arbitration strategies**	**Method Data Collection Sample size**	**Data Analysis/ Metrics**	**Main findings/results**	**Strengths/ Weaknesses**
Assess cancer detection rates, false-positive rates before arbitration, PPV for women recalled after arbitration, and the type of cancers detected with use of FFDM alone and combined with tomosynthesis	Prospective trial Quantitative design Interim analysis – phase 1	Blinded Double reading Consensus based arbitration meeting. 1. Mammography alone, 2.mammography + CAD 3.mammography + tomosynthesis 4. Synthesized mammography + Tomosynthesis	November 22, 2010, to December 31, 2011. 12631 cases 8 radiologists with 2–31 yrs. of experience in screening Images scored 1-5 One score of 2 or greater in at least one arm were discussed at arbitration before a consensus-based decision was made. Consensus-based arbitration meetings = min 2 radiologists	Analyses were based on marginal log linear models for binary data, accounting for correlated interpretations and adjusting for reader-specific performance levels by using a two-sided significance level of .0294 Cancer detection rates, false positive rates before arbitration, and PPV for patients recalled after arbitration.	**False-positive rates before arbitration were 61.1 per 1000 examinations with mammography alone and 53.1 per 1000 examinations with mammography + tomosynthesis (15% decrease, adjusted for reader; P, .001).** 5 of 8 radiologists referred proportionally more patients for arbitration with use of mammography alone than with use of mammography + tomosynthesis. **Overall number of women recalled as a result of arbitration was larger for those initially assigned a positive score at mammography + tomosynthesis (351 vs. 265 women).** However, the concordant increase in the detection of 24 additional Cancers resulted in a similar PPV for the cases ultimately recalled after arbitration (29.1% mammo alone and 28.5% + tomo)	Weakness - Only limited data about interval cancers -cannot estimate conventional absolute sensitivity or specificity. Estimate relative performance levels Potential candidates were selected on the basis of whether technical staff members and imaging systems were available to perform the additional imaging examination

19. Wolf et al (2015) *Germany* -Quality CASP criteria met						
Research question/aim	**Study design**	**Reporting/ Arbitration strategies**	**Method Data Collection Sample size**	**Data Analysis/ Metrics**	**Main findings/results**	**Strengths/ Weaknesses**
Performance of 3 collective intelligence rules ("majority", "quorum", and "weighted quorum") applied to mammography screening	Prospective Quantitative design	Majority, quorum and weighted quorum against individual radiologist performance	182 test set cases Images from 2000-2003 from 6 centres 101 radiologists randomly grouped into sizes (range: 1 to 15)	Average true and false positive rate of the no. of radiologists determined by a training set to give the quorum threshold Weighted quorum	As group size increased, all three CI rules achieve increases in true positives and decreases in false positives. Larger groups made more accurate decisions Marginal affect when group size exceeds 9 relatively small group sizes achieved performance improvements Overall decision accuracy = Weighted quorum rule slightly outperforms the quorum rule and that the quorum rule outperforms the majority rule	Strength - Large number of radiology participants – representative of diverse experience Unique, transparent system of consensus without 'over-ruling' of a group face-to-face setting. Weakness – Test set, no influence on "real-life" cases.

20. Blanks et al (1998) U*K* - Quality CASP criteria not met (No for Q6 against cohort study)						
Research question/aim	**Study design**	**Reporting/ Arbitration strategies**	**Method Data Collection Sample size**	**Data Analysis/ Metrics**	**Main findings/results**	**Strengths/ Weaknesses**
Cancer detection rates for different reading strategies. Breast screening Consensus and arbitration	Observational epidemiological study Quantitative	1. Single reading 2. Double reading (with recall if any reader suggests) 3.double reading (With recall if both readers agree, consensus) 4. Double reading (with arbitration by a third or more radiologists) 5. Double (complex)	1 April 1996 to 31 March 1997. 87 screening units	Cancer detection rate adjusting for confounding by age using Poisson regression 95% CI	Prevalent screen **Double (consensus) = 1.26 SDR** **referral rate = 6.8** **Double (arbitration) = 1.28 SDR** **Referral rate =7.3** Incident screen invasive cancer SDR - **Double (consensus) = 0.98 SDR** **Referral rate = 3.1** **Double (arbitration) = 1.10 SDR** **Referral rate =4.0** Incident screen invasive cancer SDR <15 mm Double (consensus) =1.00 Double (arbitration) =1.18	Strength - Multi –Centre study Weakness - 1yr study

21. Skaane et al (2013) *Norway* - Quality CASP criteria met						
Research question/aim	**Study design**	**Reporting/ Arbitration strategies**	**Method Data Collection Sample size**	**Data Analysis/ Metrics**	**Main findings/results**	**Strengths/ Weaknesses**
Compare double readings for FFDM (2D) and tomosynthesis (3D) during mammographic screening.	Prospective study Quantitative design	5-point rating scale for probability of cancer: 1=normal or definitely benign; 2=probably benign; 3=Indeterminate 4=probably malignant 5=malignant. Scores of 2 or greater in at least one reading arm =discussed at arbitration, with at least two radiologists Consensus-based decision for all cases with a least one rating of 2 or 3. Cases with a score of 4 or 5 were recalled and **could not be dismissed at consensus.**	22/11/10 – 31/12/11 8 Radiologists - 2– 31 years of experience (average 16 years) in screening mammography	P<0.05 Type III test -in generalised linear mixed Model (proc glimmix, v. 9.23) Heterogeneity of performance - addressed using G-side random effects	74% of mammo only cases – returned to routine recall at consensus. 26% recalled. 75% of these negative at assessment 61% of mammo +tomo – returned to routine recall at consensus. 39% recalled. 74% of these negative at assessment Pre-arbitration false-positive scores were 10.3 % mammo only and 8.5 % for 2D+ 3D (P<0.001). Recall rates were 2.9 % (365/12,621) and 3.7 % (463/12,621), respectively (P=0.005). PPV Mammo only before arbitration= 6.5% after = 24.7 % 2D+ 3D before arbitration= 10% after = 25.5 %	Strength - Scores recorded directly into the NBCSP database -results locked at the end of each reading Weakness - Unable to assess outcome of cases dismissed at arbitration – 1 yr. study

22. Hukkinen et al (2006) *Finland* - Quality CASP criteria not met						
Research question/aim	**Study design**	**Reporting/ Arbitration strategies**	**Method Data Collection Sample size**	**Data Analysis/ Metrics**	**Main findings/results**	**Strengths/ Weaknesses**
Conference consensus (the Majority considered decisive) Or Independent reading of several radiologists (the positive opinion of at least a single reader considered Decisive).	Prospective Quantitative	Double reading Conference consensus = the majority opinion in the group	1997 – 2001 200 Test cases 4 radiologists 5 -18 yrs. screening experience 2 general radiologists, 2 residents, 6 months - 4yrs. of experience in Clinical mammography.	Sensitivity/ Specificity	The greatest sensitivity of 74.5% = readings of the four best-performing readers were combined. Sensitivity very variable Sensitivity maximal when any positive opinion within a pair or a group of readers is taken into consideration. Conference reading = improved specificity	Weakness - Small number – test cases High ratio 1:4 cancers to normal cases – not representative of normal practice Actual consensus where Readers discuss discordant findings did not happen in order to avoid a situation in which one reader is overruled by another. Worked out by calculating average sensitivities

23. Matcham et al (2004) *UK* - Quality CASP criteria met						
Research question/aim	**Study design**	**Reporting/ Arbitration strategies**	**Method Data Collection Sample size**	**Data Analysis/ Metrics**	**Main findings/results**	**Strengths/ Weaknesses**
Affect of consensus on all discordant **and** concordant recalls	Retrospective and Prospective Quantitative	Consensus for all cases even if both initial readers 'recalled'	April 1997 - March 2002. 2 years prior to the start of the consensus meeting, and the 3 completed years since. 3 radiologists – 3-12 yrs. Experience 1 film reader – 4yrs experience	PPV, cancer detection rates SDR	5% of screening cases discussed at consensus meeting (n=2637) 65.6% recalled after consensus 3 interval cancers subsequently diagnosed after RR outcome following consensus – 1 true and 2 minimal signs 97 (10.7%) of the women returned to routine screening had been marked for recall by both original film readers. Consensus of all cases - Reduction in recall rates Increase in Specificity	Strength - Sufficient follow-up period to assess interval cancers and true sensitivity

24. Jenkins et al (2014) *UK* - N/A for CASP Audit						
Research question/aim	**Study design**	**Reporting/ Arbitration strategies**	**Method Data Collection Sample size**	**Data Analysis/ Metrics**	**Main findings/results**	**Strengths/ Weaknesses**
Assess differences in the film-reading histories of interval or screen detected cancers Arbitration	Audit – 3 year period	Double reading Not completely Blind reading Arbitration by 3rd reader – radiologist – not blinded has access to previous opinions	2004 -2007 4 programmes within the East Midlands Film readers – radiologists and radiographers Analogue films	Cancer detection rates, confidence intervals, and chi square Tests with Monte Carlo simulation.	Double reading= discordance in 13,279 cases (5%) underwent arbitration. 9726 (73%) were returned to routine rescreen, 3553 (27%) were recalled PPV for unanimous recall = 22.7% PPV for recall following arbitration = 8.3% 4.1% of interval cancers with no previous recall outcomes were false negatives, which was significantly lower compared to the groups where at least one reader had indicated recall (10.9%; p. 0.005). Cancers detected at the subsequent screen demonstrated no significant difference in prognosis dependent on previous film-reading history (P. 0.503).	Strengths - Robust method for identifying interval cancers.

25. Shaw et al (2009) *Ireland* - Quality CASP criteria met						
Research question/aim	**Study design**	**Reporting/ Arbitration strategies**	**Method Data Collection Sample size**	**Data Analysis/ Metrics**	**Main findings/results**	**Strengths/ Weaknesses**
Consensus review of discordant Screening mammography	Prospective Quantitative	Double reading Blinded reading Consensus panel = Three to five consultant radiologists and usually included one or both of the original readers. Recall -if any member of the Consensus panel recommended after discussion.	2000-2005 5 radiologists 3–10 years of screening experience. Two consultants who had just completed fellowship training participated for 2 years of the study period.	Sensitivity/specificity Z test (95% CI P<0.05	Discordant cases = 1.04% After consensus, 45.39% recalled **11.7% of these were cancer** Highest reader recall = could potentially increase the cancer detection rate by 0.6 per 1000 women screened but would increase the recall rate by 12.69% and the number of False-positive findings by 15.37%. Conclusion: The consensus panel identified 71 (7.33%) of 968 cancers diagnosed. Consensus review substantially reduced the number of cases recalled and was associated with a low false-negative rate. **1.1% of known cancers missed by consensus review**	Weakness - 44 (6%) cases at consensus sent to RR with no follow-up. False-negative findings was predicted by multiplying the number of patients who did not return for a follow-up visit (n -44) by the percentage of false-negative findings in patients with follow-up screening data

26. Per Skaane et al (2007) *Norway* - Quality CASP criteria met						
Research question/aim	**Study design**	**Reporting/ Arbitration strategies**	**Method Data Collection Sample size**	**Data Analysis/ Metrics**	**Main findings/results**	**Strengths/ Weaknesses**
Compare SFM and FFDM in a population-based screening program. Consensus	Prospective Quantitative	Double reading Blinded 5-point rating scale for probability of cancer: 1=normal or definitely benign; 2=probably benign; 3=indeterminate 4=probably malignant 5=malignant. Consensus meeting. Free to dismiss cases with scores no higher than 2 by one or both readers.	November 2000, and December 2001. Radiologists	Recall rate, cancer detection rate, PPV	74.1% of SFM case dismissed at consensus meeting; 68.9% of FFDM were dismissed at consensus 10.9 – 11.1% cancers missed by consensus 25-30% cancers only recalled by 1 reader	Weakness - 45-49 age group not complete follow-up? Accurate interval cancer rate reported

9. RETRIEVED ARTICLES AT END OF THE SEARCH PROCESS

A total of 715 articles were retrieved for review following publication database and grey literature searching. Excluding duplicates 601 remained. A large proportion of duplicates were the result of searching multiple databases (see PRISMA diagram). 558 of the 601-articles/grey literatures were excluded, as they did not meet the selection criteria, leaving 43 for full-text review.

10. QUALITY APPRAISAL OF RETRIEVED ARTICLES

Armstrong et al. (2011) and Booth, Papaioannou and Sutton (2012) state that scoping reviews do not usually undertake quality appraisal. However, Levac, Colquhoun and O'Brien (2010) assert that a lack of quality assessment may result in false conclusions about the issue and extent of gaps within the evidence. This view is supported by McDermott et al. (2013) for narrative reviews, which they state, can also be deemed less reliable if quality assessment is not made clear.

Critical appraisal tools provide a systematic method of pulling out pertinent information from studies and allow the reader to determine how strong the evidence is and relevance to their clinical practice. Of the retained papers quality assessment for methodological rigour was undertaken using criteria derived from the standardised Critical Appraisal Skills Programme (CASP) questions (2013) where appropriate (Appendix N). Some articles reported audit results, which were not amenable to CASP analysis. Quality appraisal of the included studies was undertaken independently by two reviewers, and in cases of disagreement, a third

reviewer was consulted with the aim of reaching consensus through discussion. 4 of the 43 papers were excluded after quality assessment, as there was insufficient evidence of reliability to warrant inclusion. No weighting or ranking of the papers finally included was undertaken.

11. CRITICAL REVIEW PROCESS

The critical review involved two distinct phases of data extraction and synthesis.

Data extraction. The main reviewer only undertook this process. This is recognised to introduce a possible level of subjectivity. Data extracted included:

- Article descriptors: author; year of publication; country where study performed;
- Study context (screening versus diagnostic);
- Sample size;
- Data analysis/metrics;
- Reporting strategy (double reading; blinded or non-blinded reading);
- Use of a test set versus prospective series of patient selection;
- Strategy utilised for discordant results;
- Readers (professions, number acting as arbitrator, years of experience, and specific training in mammogram reading);
- Strengths and weaknesses of the study.

The retrieved data was synthesised to address the primary and secondary outcomes. Details documented within the data extraction form provided a basis for the

subsequent synthesis.

Synthesis. The findings were summarised in a thematic narrative synthesis. Popay et al. (2006: 5) define this as: '*An approach to the systematic review and synthesis of findings from multiple studies that relies primarily on the use of words and text to summarize and explain the findings of the synthesis'*. Due to the heterogeneity of studies found, this method was deemed most appropriate for the present scoping review.

3.2 Chapter 3 Summary

This chapter has described the methods used and justified the need for a scoping review to address the research questions posed. It has summarised the key steps to undertaking a scoping review as advocated by Popay et al. (2006). A large volume of papers (601) were identified initially and following the screening, full text review and quality assessment this led to 26 being retained for the final review. The following chapter presents the results of the synthesis.

Chapter 4. Results

This chapter details the results of the evidence found for the scoping review on arbitration or consensus processes within mammography reporting. The chapter explores evidence on the effectiveness of different strategies utilised to resolve discordance in breast screening reports and discusses key features of the evidence, emerging themes, relationships and disparities between studies with generalizability discussed relative to UK practice.

4.1 Results of the Search

Of the 43 full text papers reviewed, 26 studies provided sufficient data for this systematic scoping review (Data extraction Table 7). The PRISMA flow chart demonstrates that 13 were excluded at full-text review, reasons for this were:

- No outcome of interest (n=7)
- No relevant data (n=6)

A further 4 papers were excluded following quality appraisal. Reasons for exclusion were:

- Historical data used as a comparator to CAD (n=1)
- Probability of the affect reported (n=1)

3 independent readers classed as the equivalent of double reading with arbitration (n=1)

No outcome in 37.4% of the cases (n=1)

4.2 Included Studies

The retained studies consisted of a mixture of designs, but all were quantitative in nature. No qualitative studies were retrieved on arbitration or consensus processes. There were eight studies (Table 7 studies 1,2,3,4,7,10,12,16) described by authors as retrospective and twelve (Table 7 studies 5,8,9,13,14,15,18,19,21,22,25,26) studies were prospective, with one (Matcham et al. 2004) a mixed design of retrospective and prospective cases. The remaining study characteristics comprised of two audits (Liston and Dall 2003 and Jenkins et al. 2014), two systematic reviews (Taylor and Potts 2008 and Dinnes et al. 2001) and one observational epidemiological study (Blanks et al. 1998). However, only five of the prospective studies (Mucci et al. 1999, Caumo et al. 2011, Ciatto et al. 2005, Shaw et al. 2009 and Duijm et al. 2004) were predominantly looking at the effect of arbitration or consensus, the remainder focused on the transition from screen film mammography to digital mammography (Skaane et al. 2007), comparison of current reading protocols to CAD assisted reading (Khoo et al. 2005), impact on the number of readers (Hukkinen et al. 2006, Wolf et al. 2015) and comparison of conventional FFDM with tomosynthesis (Lang et al. 2016, Skaane et al. 2013a and Skaane et al. 2013b). Two systematic reviews

(Taylor and Potts 2008 and Dinnes et al. 2001) were incorporated but their primary remit was comparison of reading strategies i.e. double reading with single reading, and single reading with and without CAD.

4.3 Data Extraction of Study Features

Table 7 summarises the pertinent characteristics of the included studies. The features extracted from each publication were authors, year of publication, country, study design, research aim, sample size, characteristics of the participants, duration of the study, reporting/arbitration strategies, data analysis/metrics, main findings, and strengths/weaknesses. Significant differences amongst studies were found for all characteristics considered.

4.3.1 Publication Date

Publication dates ranged from 1998 to 2016, with two studies (Klompenhouwer et al. 2015a and Klompenhouwer et al. 2015b) both within 2015 consider the same cohort of patients, but are published as two separate sub-studies, although arbitration was the primary focus of both. The number of studies published per year varied from 0-4, with the peak number in 2009 (n=4). It is therefore notable that a number of studies have been undertaken prior to the start of the digital transition in 2006.

4.3.2 Country of Publication

Publications have predominantly been from the UK (n=11) with the last publication being a 2014 audit; prior UK studies relate to 2009 or earlier. The remaining

publications were from the Netherlands (n=4), Norway (n=4) and Italy (n=2) with one publication from each of Australia, Finland, Sweden, Spain and Germany. Again, two of the Netherlands studies (Klompenhouwer et al. 2015b and Klompenhouwer et al. 2015a) represent the same cohort of patients and study period.

4.3.3 Characteristics of the Readers

There was variability in both the experience and cohort of professionals' undertaking the reporting process. In those studies, where information was provided, radiologists' experience is given as a range, which varied from 15 months screening experience to more than 30 years. Five studies incorporated radiographers as film readers but this would be consistent with the UK only, who utilise this cohort of staff as a reporter and only one of these papers was a fairly recent (2014) audit. James and Cornford (2009) represent the latest research study, which incorporated one radiographer as a film reader. Radiographer reporting is now commonplace in the UK with a number of units utilising double radiographer reporting following the NDROR trial in 2012 (Bennett et al. 2012). Therefore, no recent arbitration or consensus studies were found reflecting this change in reporting personnel. Two studies (Cornford et al. 2005 and James and Cornford, 2009) incorporated research fellows within the reporting group, one study (Shaw et al. 2009) included two consultants who had just completed fellowship training, a further study (Cawson et al. 2009) utilised a senior radiology trainee with 6 months training, and one study (Hukkinen et al. 2006) included 2 general radiologists and 2 residents (equivalent to a UK House Officer). Internationally, specialist and general radiologists are representative of the

workforce reporting screening mammography, which is disparate to current UK practice.

4.3.4 Population/Sample Size

All studies were within population-based national or regional screening programmes. Sample sizes within the studies ranged from 182 test set cases to a retrospective review of 1,033,870 prevalent and incident screens. Study duration varied greatly dependent on study design ranging from a 4 month prospective study (Caumo et al. 2011) to a 9 year retrospective study (Hofvind et al. 2009).

4.3.5 Test Sets

Five studies utilised test sets (Cawson et al. 2009, James and Cornford 2009, Groenewoud et al. 2007, Wolf et al. 2015 and Hukkinen et al. 2006). A weakness identified by the authors James and Cornford (2009) was that although 240 real arbitration cases were utilised to test the effectiveness of CAD as an arbiter, using a small number of test cases meant there was only a small number of cancers (18) in the series. The other four studies circumvented this by weighting the test sets with relatively high ratios of cancers to normal cases; however, this therefore does not reflect the normal screening scenario. In clinical practice readers need to report relatively large quantities of films before a positive cancer case is read.

4.3.6 Double Reporting - Blinded and Non-blinded

Not all studies detailed the percentage of cases in which double reading produced

discordant results, but from the data available there was a large variation ranging from 0.5% (Duijm et al. 2004) to 57.2 % (Klompenhouwer et al. 2015a) of cases. Klompenhouwer et al (2015a) demonstrated that there was a significant difference in the number of discrepant reads dependent on whether the double reading was performed blinded (57.2%) vs. non-blinded (29.1%). This raises the question as to whether some of the blinded-reading studies are truly blinded. The UK authors acknowledged that readers were not completely blinded as only one set of assessment paperwork is commonly utilised. Upon completion of the reading batch, the 2nd reader is required to amalgamate the paperwork and hence will be aware of the 1st reader report and any discrepancies. This therefore presents the potential for the 2nd reader to review the case and change their opinion. This may subsequently affect an individual's cancer detection rates.

4.4 Arbitration Studies

Although twelve studies (1,2,4,5,6,7,8,9,10,13,18,24) mentioned arbitration within the reporting process only five studies (1,2,5,8,9) were specifically looking at the effectiveness of the arbitration process, with one study reporting the affect of CAD acting as an arbiter (James and Cornford 2009). Within other studies arbitration occurred as part of normal clinical practice but the main focus of the studies was evaluating CAD performance, tomosynthesis, film reading histories or assessing the mammographic and pathological features of screen-detected and interval cancers.

4.4.1 Effect of Arbitration on Recall Rates

In those studies, that provided the information, the final decision of the arbitrator resulted in wide variation as to whether cases were ultimately recalled or discharged back to routine screening. In the Jenkins et al (2014) study 27% of cases were recalled to assessment following arbitration but in the Cornford et al (2005) study 50% of cases were recalled. There was a comparable variation in cases returned to routine screening following arbitration ranging from 50% (Cornford et al 2005) to 72.4% of cases in the Caumo et al (2011) study.

Overall, studies reported that compared to highest reader recall (non-arbitration), arbitration resulted in significant reductions in recall rates, with relative decreases in the range of 17.8% (Klompenhouwer et al 2015b) to 40.9% (Caumo et al. 2011). Although Caumo et al (2011) report one of the highest reductions in recall rates by arbitration, the results must be interpreted with caution as this study was conducted over a short (4-month) period, with a single experienced (>30yrs) individual arbiter. All cases were recalled to assessment irrespective of the arbitrator's decision, and therefore there was no direct impact on clinical care. Variability in reducing recalls is also confirmed by Liston and Dall (2003) reporting findings from a seven-year audit.

Klompenhouwer et al (2015a) assert the reporting strategy (blinded vs. non-blinded) also has a significant effect on sensitivity following arbitration. Although no effect on sensitivity was reported at non-blinded reading, blinded reading with arbitration demonstrated a statistically significant decrease in sensitivity (83.2 to 76.0 %, p =

0.013). With such variation in recall rates the PPV of assessment cases following arbitration is also unpredictable with low PPV's of 8.3% (Jenkins et al. 2014) to 31.2 % (Klompenhouwer et al. 2015a) reported.

There is disparity between the studies regarding the effect of arbitration on cancer detection rates. Klompenhouwer et al (2015b) declared an overall decrease, albeit it small (0.1-0.2%) and not statistically significant, whilst the systematic review by Taylor and Potts (2008) stated double reading with arbitration increased cancer detection rates. Dinnes et al (2001: 458) systematic review affirmed there was '*insufficient evidence to detect any pattern in cancer detection based on recall policy*'.

The scoping review highlighted that within the five studies (1,2,5,8,9) specifically looking at the effectiveness of the arbitration process, two studies were retrospective analysis (Klompenhouwer et al. 2015a and Klompenhouwer et al. 2015b) and although Caumo et al (2011) was a prospective design, assessment was performed irrespective of the arbitration decision. Methodological weaknesses were also identified in the Caumo et al (2011) study which was undertaken over a short 4-month period, and utilising only 1 radiologist as the arbitrator who had extensive experience (>30yrs). The James and Cornford (2009) study was unique in using CAD as a means of assisting interpretation where discrepant lesions had already been identified by one of the initial readers. The decision of CAD as the arbitrator was retrospectively compared with the decision of the original third independent reader and hence like the Caumo and Klompenhouwer studies the results can only

hypothesise the effect of the arbiter because the outcome did not affect real care.

Therefore, it is difficult to confirm whether the decision-making of returning cases to routine recall would be the same when it has a direct immediate impact on clinical practice. Subsequently only two studies (Mucci et al. 1999 and Ciatto et al. 2005) provide this information. Mucci et al (1999) were establishing cost savings by arbitration and the authors acknowledged that non-blinded reading occurred and the second reader may therefore have been influenced by the first reader's decision. Ciatto et al (2005) were also investigating the effectiveness of arbitration but did not use continuous cases, this was limited to periods when radiologists were available to perform a third read and follow-up data for 58% of the cases in which arbitration concluded a negative outcome were not available. Therefore, the effect of the arbitration process and subsequently cancer detection rates could only be estimated.

Although CAD and Tomosynthesis were not a primary focus of this review, it was noteworthy that the evaluation of CAD as an arbiter although more sensitive than an independent 3rd reader, showed a statistically significant increase in normal cases being recalled to assessment with a relative increase of 10% (James and Cornford 2009). Tomosynthesis studies (Skaane et al. 2013a, Skaane et al. 2013b and Lang et al. 2016) retrieved via this review also reported that there was a statistically significant difference in the overall number of women recalled as a result of arbitration in the cohort undergoing conventional FFDM + tomosynthesis versus FFDM alone. Two studies (Cornford et al. 2005 and James and Cornford, 2009)

provide information that arbitration cases account for 20-22% of assessment cases in the UK (range 20-22%), which represents a significant proportion. Therefore, although emergent technologies may improve the cancer detection rates, consideration needs to be given to the impact of extra cases requiring assessment.

4.4.2 Consensus Studies

Five papers (15,21,23,25,26) mentioned consensus as the method of resolving discordant cases with only two of the studies (Shaw et al. 2009 and Matcham et al 2004) specifically looking at the effectiveness of the consensus process. The three remaining studies were evaluating CAD and tomosynthesis. Therefore, limited data was available on recall rates to assessment following consensus with a range 31.1% (Skaane et al. 2007) to 65.6% reported (Matcham et al. 2004), and 68.9% - 34.4% of cases being returned to routine recall. The high number of cases returned to routine recall in the Norwegian study (Skaane et al. 2007) relates to the scoring system utilised where a score of 2 (defined as probably benign) or greater is referred for consensus discussion. As per the arbitration studies, more women were allocated to consensus when mammograms were read with CAD (13.8% vs. 10.5%) to non-CAD reading (Khoo et al. 2005).

4.4.3 Mixed Studies/Reviews

Within eight of the nine mixed studies (3,11,12,14,16,17,19,20,22) it is not possible to differentiate the effect of arbitration versus consensus as the processes are either integrated in the discussion, or both are undertaken within the decision making

strategy i.e. mutual consensus between the two readers with persistent discordant case being reviewed by an arbitration panel. The Duijm et al (2004) study reports that this strategy resulted in 45% of cases being resolved by mutual discussion and 55% still requiring arbitration by a panel, with 48.6% of the cases subsequently recalled. The panel recalled if at least one arbitration member considered it necessary. However, there is no information provided on the agreement levels between the 3 panel arbitrators, this would therefore raise the question; would recall rates/detection rates have been different with a majority decision? Recalling based on one member's decision may have resulted in higher recall rates comparative to a majority decision and the subsequent effect on PPV remains unknown.

Minimal pertinent information can be extracted from the (Posso et al. 2016) cost and health related outcome study (comparing single vs. double reading) other than discordant reading occurred in 4.5% of cases with 98.1% resolved by consensus and 1.9% still requiring arbitration. No interval cancer rates are provided and therefore results are not conclusive. Groenewoud et al (2007) although a paper primarily concerned with cost effectiveness of different reporting strategies stated that referral rates were highest with decision-making by consensus (73.8%) compared to arbitration (52.7%). However, this was an experimental study with test cases and therefore not reflective of clinical practice. Published regional data was utilised to estimate the distribution of discrepant readings and there was an assumption that each referral of a case would lead to a diagnosis of cancer, which is not a

representative distribution within an assessment clinic. Conversely, Blanks et al (1998) studied cancer detection rates for a variety of reading strategies and concluded that although consensus had a lower recall rate, the Standardised Detection Ratio (SDR) was higher for double reading with arbitration compared to double reading and consensus for both prevalent and incident screens. Also, for incident screens the SDR for small (<15mm) invasive cancers was also higher (Double consensus =1.00 vs. Double arbitration =1.18). This is noteworthy as the clinical perspective indicates a move to consensus over arbitration.

Hukkinen et al (2006) was a small study involving 200 cases and although describing independent reading and conference consensus (the majority considered decisive) stated that they avoided readers discussing discordant cases to prevent the situation of one reader being overruled by another. Consensus was calculated by average sensitivities and this achieved maximum results (75.4% sensitivity) when combining the readings of the four best performers. This is similar in principle to the unique Collective Intelligence (CI) study (Wolf et al. 2015) that utilised a majority, quorum and weighted quorum rule tested against an individual radiologists performance. In accordance with Hukkinen et al (2006) as group size increased all three CI rules achieved increases in true positives and decreases in false positives. Larger groups were declared to make more accurate decisions, but relatively small group sizes achieved performance improvements. However, this was again a test set scenario with no influence on real-life cases.

A further variation in recall policy was discussed by Hofvind et al (2009) and Matcham et al (2004) who performed consensus on all recalls (concordant and discordant) resulting in 17.9% and 10.7% of the concordant readings to recall being over-ridden at consensus.

4.4.4 Discordant Cancers

Discordant cases that were subsequently histologically proven to be a cancer ranged from 2% (Caumo et al. 2011) to 23.6% (Hofvind et al. 2009), but with a very short (4 month) study period the results of Caumo et al (2011) need to be interpreted with caution.

4.4.5 Follow-Up/False Negative Cases

Regardless of the strategy used cancer cases were incorrectly dismissed to routine recall by both processes. Only twelve studies provided information regarding interval cancers; the length of follow-up was variable ranging from four months to seven years, and as a full screening interval (2 or 3years dependent upon country) was not complete prior to the reporting of some studies the true effect of cases returned to routine screening is unknown. Ciatto et al (2005) reported that 0.64% of dismissed arbitration cases were false negative but as discussed previously this is an incomplete data set with only 42% of cases followed up. Shaw et al (2009) and Duijm et al (2004) report fairly low rates of cancer cases dismissed at consensus 1.1% and 3% respectively. Interestingly, Jenkins et al (2014) showed 4.1% of false negative interval cancers were double reported as normal, which was significantly

lower compared to cases where at least one reader had indicated recall (10.9%; p< 0.005). Following a review of consensus false–negative cases, Shaw et al (2009) report a change in practice recalling a much higher proportion of discordant cases when microcalcification is the mammographic abnormality.

4.4.6 Tumour/Mammographic Characteristics Of Discrepant Cases.

Klompenhouwer et al (2015a) described no difference in the proportion of DCIS, smaller tumours, lymph node involvement or advanced tumours between screen-detected cancers and those missed at arbitration. Conversely Cornford et al (2005) indicate cancers detected following arbitration were smaller (p<0.045), a finding also supported by Cawson et al (2009). Lobular cancers which are often mammographically difficult to detect were reported to be more common in the arbitration group, albeit of borderline significance (Cornford et al. 2005: 1186). The mammographic features were '*more likely to present as parenchymal distortions (p<0.001), and less likely to be detected in fatty breasts (p<0.01)*'.

4.5 Chapter 4 Summary

This chapter reported the synthesised data collected on arbitration and consensus processes. It was found that there was a limited body of evidence relating to either processes and in particular a lack of prospective studies to determine their effectiveness in real-life clinical settings. Methodological weaknesses were identified in some studies, and predominantly the lack of complete follow-up or reporting of true interval cancers compromises the ability to conclude the

effectiveness of the processes. The following chapter reports the themes arising from the data synthesis, discusses the limitations of the review and proposes recommendations for further research.

Chapter 5. Discussion

The aim of this systematic scoping review was to establish what evidence there is to inform models of arbitration or consensus in mammography reporting, with a secondary aim of identifying gaps in the evidence base and recommendations for further research if required. The scoping review has revealed a dearth of literature relating to process, in particular a lack of prospective studies demonstrating effectiveness of different processes in relation to outcomes (recall rates, cancer detection rate, PPV and programme sensitivity/specificity). There is considerable variance in the processes used, a lack of guidance and a number of key areas where no evidence was retrieved. This chapter discusses the details of the emergent themes.

5.1 Lack of Guidelines

The NHSBSP guidance (2011: 7) specifies that services unable to achieve the minimum standards for recall rates (prevalent and incident) must '*carry out arbitration as a matter of routine*'. Review of the guidance revealed no information specifically relating to consensus. This was also evident from the literature searching as no guidelines were identified on how to attain consensus if there is a discrepancy between the two readers. The lack of guidance has resulted in breast unit's implementing a variance in practice. These differing work practices were not just evident across the UK but internationally, and therefore comparisons of outcomes are problematic.

5.2 Variations in Practice

The scoping review revealed that there are inconsistences not only in which method is used to resolve the discordant cases, but also within the structure and scheduling of the processes. It is apparent from personal contact with experts that scheduling of arbitration or consensus in the UK ranges from ad-hoc impromptu arbitration to scheduled/timetabled consensus meetings. The former reflects the practitioners experience and the lack of defined periods to review cases may not represent best practice.

5.2.1 Different Definitions

Definitions of consensus and arbitration are not clear-cut. This was evident from the studies retrieved via the literature searches as well as from direct contact with clinicians. The two terms are used interchangeably and often confusing with some studies reporting 'arbitration by an individual', others 'arbitration by a panel', and 'consensus based arbitration' meetings. This confusion has also been experienced from the practitioner perspective at regional meetings where individuals debate whether the process they are utilising is technically classed as arbitration or consensus. The literature review also highlighted a lack of consistent terminology regarding how the reading and arbitration strategy was undertaken. Some utilised the term independent and others blinded to indicate that the second reader was unaware of the first reader's decision; or the arbitrator was unaware of the reason for recall. The lack of clear definitions makes it not only difficult to review the

literature and synthesise the findings, but it also adds to confusion in a clinical setting when discussing processes with no clear delineations.

5.2.2 Different Approaches

Within consensus and arbitration processes a range of approaches were used to reach the final decision. Scenarios ranged from a 3rd independent reader who made the final decision to an independent review by a 3rd reader but with the majority opinion acted upon. In other studies, cases were sent to arbitration only if the first readers could not achieve consensus; a panel (that may or may not incorporate the original reporters) then arbitrated persistent discordant cases. The decision process for referring to assessment was also diverse ranging from a majority decision to acceptance if at least one member specified recall. Inconsistencies were also evident in whether the arbitration was performed blinded, or whether the reason for the recall was made apparent and the role of arbitration was to decide on the action to be taken on an identified lesion. The latter reflects the clinical practice experienced by the author.

The literature review retrieved only one study (Hofvind et al 2009) that reported consensus of all recalls (concordant and discordant). Contact with professional associates identified one unit with a previously high recall rate, which discussed all recalls at consensus meetings and frequently overrode the decision of both original film readers at the consensus review. Conversely, some professionals also reported the scenario of sending all prevalent recalls (concordant and discordant) for

discussion at a consensus meeting. The author's clinical perception is that centres are moving to group consensus rather than a 3rd reader arbiter, but the number of units adopting this practice remains unknown at present. This change in practice may relate to group consensus offering an opportunity for educational learning from cases, a perception that groups will miss fewer cancers or the fact that responsibility for decision-making should not lie with a particular individual. It also raises the possibility that fear of litigation is an additional factor.

5.2.3 Different Scoring / Classification

A further area making international comparisons difficult was the disparity in scoring systems used to grade the mammographic images. This was dependent upon the country of the study. Differences in scoring systems may affect the perceived size of disparity and therefore which cases are sent to arbitration/consensus. In the Netherlands, studies utilised the Breast Imaging Reporting and Data System (BIRADS) produced by the American College of Radiologists (ACR), which works on a 0-6 scale. The Norwegian Breast Cancer Screening Program (NBCSP) utilised a unique system of scoring of 1-5. This was essentially like the Australian (National Breast Cancer Centre) (Cawson et al. 2009) and Swedish (Lang et al. 2016) studies which report a 1-5 category.

The UK Royal College of Radiologists (RCR) Breast Group (RCRBG) system also works on a 1-5 scale but unlike the BIRADS categories does not provide a probability (%) of cancer and the short-term 6-month follow-up (category 3) is not utilised. The

NHSBSP uses an IT system (National Breast Screening Service - NBSS), which equates to the RCRBG system, but assessment paperwork is documented with codes rather than numbers as demonstrated in Table 8 below.

Table 8 A summary of the differences between the classification systems

UK RCR category	UK management	NBSS Category	NHSBSP Management	BIRADS category	BIRADS Management	BIRADS Likelihood of malignancy (%)
	N/A	N/A	N/A	0 Incomplete—need additional imaging evaluation and/or prior mammogram for comparison	Recall for additional imaging and/or comparison with prior examination(s)	N/A
1. Normal	Routine screening	1. N-normal	Routine recall	1 Negative	Routine screening	Essentially 0% likelihood of malignancy
2. Benign	Routine screening	2. B-Benign	Routine recall	2 Benign	Routine screening	Essentially 0% likelihood of malignancy
3. Indeterminate/ probably benign findings	Further investigation is indicated.	3. U - Uncertain	Recall	3 Probably benign	Short-interval (6-month) follow-up or continued surveillance imaging	>0% but ≤2% likelihood of malignancy
4. Findings suspicious of malignancy	There is a moderate risk of malignancy. Further investigation is indicated.	4. S – Suspicious	Recall	4a. Low suspicion for malignancy 4b. moderate suspicion for malignancy 4c. high suspicion for malignancy	Tissue diagnosis	> 2% to ≤10% > 10% to ≤50% > 2% to <95%
5. Findings highly suspicious of malignancy	There is a high risk of malignancy. Further investigation is indicated.	5. M - Malignant	Recall	5. Highly suggestive of malignancy	Tissue diagnosis	≥95% likelihood of malignancy
	N/A			6. Known biopsy - proven malignancy	Surgical excision when clinically appropriate	N/A

Although Taylor et al. (2011) demonstrated that the UK RCR system could be transposed to the BIRADS system; the comparison of data with differing systems may serve to add confusion. The Norwegian studies (Skaane et al. 2013a and Hofvind et al. 2009) describe a cumulative scoring process that determined which cases are sent to arbitration/consensus and subsequent management. An initial score of 2 or higher by either reader resulted in consensus, and a score of 3 or higher could not be returned to routine screening without agreement from the original reporter. Cases scoring 4 or 5 would not be dismissed.

5.2.4 Different Recall Rates/Reporting Professional

European guidelines recommend lower recall rates comparative to the UK programme and one of the strategies endorsed by Duijm et al. (2004) was the use of the highest reader recall. This process requires that if any reader deems the mammogram abnormal, the case be automatically recalled for further assessment. However, this strategy is based on a remarkably low Dutch recall rate of 0.9% and would, therefore, have significant clinical implications in the UK setting with much higher recall rates.

The literature revealed that in some countries, general radiologists were reporting mammography, as specialisation in screening is not mandatory. Expert contact made during the review provided details from a 2015 American workshop (National Academies of Sciences, Engineering, and Medicine 2015). This report summarised that in the USA some facilities have general radiologists reading all their

mammograms, others have radiologists who specialise in breast imaging, and hybrid situations exist where general radiologists perform the initial reading, with breast imaging specialists undertaking workups and biopsies. This is disparate to UK practice where current quality standards include that all film readers (Radiologists and Radiographers) in the NHSBSP must complete a recognised course of study, have 2 years' regular film reading experience, read a minimum of 5,000 mammograms per year and participate annually in the Personal Performance in Mammographic Screening (PERFORMS™) reporting test.

Overall, screening outcome is influenced by many interrelated factors and the disparities in screening interval; classifications, reading strategies and reader performance make comparisons problematic.

5.3 Lack of Evidence

From the literature reviewed it is apparent that the published data relating to arbitration and consensus is limited and great variability exists in how final decisions are made. No research was found comparing the accuracy of an independent 3rd reader (arbitrator) versus consensus (group/panel review) of discordant cases.

There was a supposition from some of the literature that fewer cancers will be missed by panel consensus compared to single reader arbitration. However, no evidence was found to support this. This is notable, as from a practitioner perception based on QA visits, and peer discussion I am very confident that the

majority of unit's based in one large region have moved to consensus processes in favour of arbitration. This view was also supported by e-mail correspondence from expert contact outside of the author's local region. The rationale for this change in practice is unclear. UK studies have elucidated that this might reflect the change in professional skill mix within the UK breast reporting system, but as many of the studies related to European countries where only radiologists would constitute consensus panels, there is a lack of evidence.

The short time interval from reporting to review at consensus provides the opportunity for individuals to evaluate trends in personal missed/misdiagnosed cases. Individuals are more likely to remember the circumstances and rationale for recalling/not recalling and hence the difference between a perception error and a decision-making error is evident. However, the dynamics within the consensus meetings can be a significant factor affecting the final decision as all individuals' opinions should have an equal weighting and everyone must feel able to voice if they strongly disagree with the decision being made.

No evidence was retrieved on how consensus meetings could be optimally structured. In particular, no studies were retrieved which examined the dynamics within breast consensus meetings and no research evaluated the complexities within a hierarchical structure undertaking decision making on discordant breast cases. The collective intelligence study (Wolf et al. 2015) provided an interesting perspective as it removed the hierarchy and difficulties associated with group decision-making.

However, this model required multiple reads to evaluate a mammogram, and this may be problematic if units are struggling to achieve screen to results within a two-week period.

From a practitioner perspective, it is imperative that consensus meetings are scheduled to allow film readers to attend on a regular basis if they are to achieve improved performance. However, the ultimate outcome post assessment or biopsy is the only way to confirm whether recall was justified and individuals will still need to audit their results. Equally important may be that, if a particular Consultant (Radiologist or Radiographer) initially reports the case as normal; they will not be the best person to perform the work-up at assessment, as they either did not perceive the abnormality or may be predisposed to report a benign finding.

Interestingly, the Blanks et al. (1998) study concluded that although consensus lowered the recall rate, the SDR was higher for double reading with arbitration for both prevalent and incident screens and smaller cancers (<15mm). The audit undertaken by Jenkins et al. (2014) identified that no excess of interval cancers classified as uncertain or suspicious were returned to routine recall after arbitration. However, a significant message from this and the Hofvind et al. (2009) study was that the interval cancer rate was substantially higher in cases that had undergone arbitration or consensus relative to the rate among concordant negative screenings. Jenkins et al (2014) report that 19.4% of interval cancers categorised as uncertain and suspicious were not initially called by any reader compared to 36.1% that had

been recalled by at least one film reader ($p < 0.001$). This raises the question of whether arbitration or consensus could be refined to aid earlier detection in such cases.

Although studies have investigated the fatigue associated with screen reading, no studies were found which evaluated the impact of performing arbitration or consensus at a particular time of day, the day of the week, the duration of consensus meetings, or the impact of the immediate prior task on decision making.

The scoping review emphasised that a number of the studies (five) utilised test cases. Although this provides a means of evaluating a reasonable number of discrepant cancer cases in a short period, readers will always be aware of the test situation and that test sets will be loaded with a higher proportion of cancers to normal cases which may affect performance levels. The immediate feedback provided by test sets serves as a valuable learning process however; the correlation of performance with real-life clinical outcomes remains an area for further research.

Overall, either the short follow-up period, lack of complete data, absence of reporting of true interval cancers versus false negatives and the retrospective nature of many studies means there is insufficient evidence to assess the effectiveness of one strategy versus the other.

5.4 Emerging Technologies

Several studies included an assessment of CAD or DBT. Although an evaluation of these was beyond the scope of the review it was notable that both technologies impacted on the number of arbitration cases and subsequent recalls.

The 2013(b) Skaane et al. study demonstrated that although 62% of radiologists referred fewer patients for arbitration with the use of FFDM and tomosynthesis the overall number of women recalled after arbitration was larger for this cohort (351 versus 265), which was also supported by Lang et al. (2016) and Skaane et al. (2013a). The authors hypothesised that the higher recall rate was a result of reader bias in favour of mammography at the arbitration meetings, which may reduce as the confidence in a new procedure develops. A systematic review retrieved in the final search in Scopus (4th June 2016) concludes that, compared to FFDM alone, DBT with FFDM increases invasive cancer detection rates and may reduce false negative recalls (Hodgson et al. 2016). An exception within the studies reviewed was the Norwegian OTST study that reported higher recall and false-positive rates after arbitration for DBT and FFDM, but potential biases were acknowledged that might explain this.

An important factor related to the use of these new technologies is that they may detect more cancers and hence produce more recalls. Therefore, the role of arbitration and consensus will be paramount in reducing false positives, as resources within assessment clinics are already limited in some services. Klompenhouwer et al

(2015a: 2828) recently suggested that CAD might be utilised as an arbitrator when '*no other method of consensus or arbitration is available'*. Within this review, the CAD studies identified were primarily concerned with aiding detection of lesions rather than assisting the decision making process. The James and Cornford (2009) study was unique in investigating the potential of CAD as an arbitrator, but this study as with others indicated that CAD produced too many false prompts. However, these studies were undertaken in 2009 or earlier, and CAD systems may well have evolved so that scales of suspicion could be of practical clinical use.

As DBT is showing increased cancer detection rates across all breast densities, if implemented at the screening stage this may represent the opportunity to assess whether second reading of the BIRADS category one cohort (almost entirely fat, glandular tissue is less than 25%) is justified. Single reading would be an entirely new concept to many readers within the current workforce and assurance of quality performance would be a pre-requisite.

5.5 Cost Analyses

Economic analyses were not the main focus of this review. However, some retained papers were primarily concerned with the cost of differing reporting strategies.

5.5.1 Length of Read

Reading times in the CAD cost analysis (Khoo et al., 2005) were assumed to be a mean 25 seconds per case for the initial report, plus arbitration was assigned 2.2

minutes of a radiologist's time. Essentially arbitration is the 3rd read and therefore should not take a disproportionate amount of time to report. However, the difference above acknowledges that consensus discussion is likely to take longer. Any future cost analysis comparing these two strategies would need to determine accurate resource use (personnel grade and time).

5.5.2 Impact of Changes In Practice

The costings reported in the Mucci et al (1999) study related to FNA. NCB or Vacuum-Assisted Biopsy (VAB), which is more expensive, has largely replaced this and therefore the Mucci et al (1999) study is not comparable to current UK practice. Methodological weaknesses were also identified in the Groenewoud et al (2007) study which was undertaken in an experimental setting and therefore not reflective of daily practice, with an assumption that all cases recalled would be cancer. This is disproportionate to a real-life assessment clinic.

The overall cost of assessment is multifaceted. It depends on the number of cases initially recalled and the proportion that are subsequently positive. A high recall rate would indicate that a greater percentage of cases would be negative or benign. As discussed in chapter 2, during the assessment, the workup of cases can follow many pathways. The resources (staff time and consumables) required for these may be vastly different. It is, therefore, complicated to cost an assessment episode (particularly false-positives with an unknown work-up variable). Given the current workforce shortages the use of a 3rd reader arbiter versus a consensus meeting

involving a group of individuals is an important consideration in terms of available skills as well as costs. Costs are usually equal in terms of additional imaging. But, depending on how departments run; there is the potential for breast care nursing and administrative costs to vary between assessment clinics.

The most recent study of costs and health-related outcomes (Posso et al 2016), evaluated double versus single reading of mammograms. From a UK perspective the costings are not directly transferable as an FNAC is costed as more expensive (€141.8) than a biopsy (€131.7). The converse is true in UK practice and as previously discussed; NCB is the recommended sampling method. The authors acknowledge that the study was performed without information about interval cancers hence the results are not conclusive. A cost-effectiveness evaluation would be required to confirm '*Whether double reading is still necessary at digital screening mammography*' (Posso et al 2016: 10). From the limited studies included in this scoping review, no conclusions can be drawn regarding cost effectiveness.

However, a consequence of double reporting is the associated delay in delivery of screening results. Individual units will differ in the delay before the 2nd read is performed. It can be difficult if there is a shortage of film readers or the infrastructure does not provide backfill-reporting sessions during periods of annual leave (the delay could be days) putting pressure on services to achieve the screen to assessment target (within three weeks). Although the strategy employed by Skaane et al (2013b) may represent low thresholds for arbitration, it does raise the question

of whether there is value in second reading cases classified as a 5 (malignant) by the first reader. It could be that the 2nd reader may detect further foci of disease or contralateral areas of concern, but as the person leading the assessment is obliged to review the case as a whole this provides the opportunity for review and workup of any further areas of concern considered necessary. A single read of these cases may represent efficiency savings in resources and reduce patient anxiety as a recall to assessment could be instigated earlier. It may be useful to establish, in cases classified as malignant by the 1st reader, what the pick-up rate of further disease is by having a 2nd read. It is uncertain how many cases in the cohort graded as suspicious (4) or malignant (5) by one reader are overridden at arbitration or consensus; or whether all cases lie within the uncertain (or BIRADS 0) category. This cohort has been shown to have a low PPV at recall and, if prevalent recall rates are high in some units, this may be the category that requires rigorous scrutiny and evaluation of whether cases are being over recalled. This may also be the category that benefits from review of concordant as well as discordant recalls as the level of suspicion may be low for both film readers.

5.6 Discussion of Method and Limitations

5.6.1 Methods

The strength of this scoping review is the systematic approach adopted and the range of evidence that was identified to provide a comprehensive overview of the '*volume, nature, and characteristics of primary research*' in the pre-specified areas (Arksey and O'Malley, 2005: 6). Seven databases were searched, some with multiple searches strategies to incorporate arbitration and consensus processes. A large volume of abstracts was reviewed (601), with 43 full text papers examined. Although only 26 studies met the final inclusion criteria this reflects the lack of evidence within this field.

With a short time frame consideration had to be given to what was practically achievable. Therefore, the scope was limited to the processes while endeavouring to ensure that the ability to answer the research question was not compromised. Use of two reviewers helped to guard against bias in application of inclusion and exclusion criteria, and search strategies to aid reproducibility. However, it is recognised that a single reviewer performed extraction and synthesis of pertinent data. These stages therefore, remain subjective and it is possible that different interpretations of the same set of studies might occur if undertaken by another individual. Nevertheless, a clinician undertook the data extraction and therefore the complexities of the clinical components of a study may be better interpreted.

Efforts were made to undertake a comprehensive review. For the search strategy to identify a manageable number of papers, the scope of the review was limited to the effectiveness of the processes, exclusive of cost. Although the search terms selected were wide-ranging, other terms may have identified further articles. It is also possible that not all studies in the published or grey literature were sourced. Searches were not designed to incorporate a comparison of single reading versus double reading. Attributes of the personnel undertaking arbitration was also considered beyond the scope of the review and therefore no conclusions can be drawn on how an individual's experience, decision-making skills, audit or reflective learning, affect the processes.

5.6.2 Limitations

The QUADAS tool was developed as a means of assessing the quality of diagnostic accuracy studies incorporated into systematic reviews, (Whiting et al 2003a); but results may be biased if aggregated data have not been individually evaluated for quality. Whiting, Harbord and Kleijnen (2005b) also identify that a major problem in quality scoring is the lack of objectivity in defining the weighting criteria. As different models utilise varying principles and weightings, combined scores can be diverse with no indication of which represents maximum dependability. Whiting, Harbord and Kleijnen (2005:7) conclude, '*Quality scores should not be incorporated into diagnostic systematic reviews*'. In 2016 Whiting et al (2016c) reported, 'A new tool to assess the risk of bias in systematic reviews' (ROBIS) as '*flaws or limitations in the design or conduct of a review have the potential to bias results'* (Whiting et al

2016c: 226)

Although quality assessment is not usually undertaken in scoping reviews, it was considered that critical appraisal of included studies could help identify any flaws that might bias the findings of the report as a whole. To make quality assessment feasible, critical appraisal was limited to whether diagnostic studies met the CASP criteria, but there was no scoring or ranking assigned.

Because searches were limited to articles published in English this may have excluded some relevant studies. As the initial searches were undertaken less than five months ago, it was not deemed necessary to re-run these on all databases. The database (Scopus) that had retrieved the greatest number of relevant papers was selected and searches re-run on the 4th June 2016. It is acknowledged that this strategy may not have retrieved all recent papers.

A decision was made to include the Ciatto et al. (2005) study although methodological weaknesses were identified as incomplete data was reported. As it represents one of the few prospective studies primarily reporting the effectiveness of arbitration, it was considered important to retain. Study authors could have been contacted to request more information, but as the research was undertaken in 2005 it was decided that if a full data set was available a further study would have been published. Nevertheless, the judgement to record the strengths and weaknesses of studies in the data extraction table provided an indication of the robustness and limitations of the results.

Because of the time constraints, working practices could only be described from a few specific professional contacts. It was beyond the scope of this study to gain a national perspective. Further evidence from a larger and wider ranging number of centres within the UK could have identified a wider range of strategies.

5.7 Conclusions

The scoping review presented in this thesis has explored the available evidence on models of arbitration or consensus in mammography reporting. A limited number of studies were identified which have assessed the effectiveness of either strategy and, within these; there was heterogeneity in study design, definitions and outcomes. The review has identified a lack of guidance and underpinning evidence to inform how best to use arbitration or consensus to resolve discordant reads, and that no current system correctly recalls all discordant cancer cases.

The purpose of the review was not to criticise the methods that breast units have employed to resolve discordant reports but to provide an understanding of the processes that are used. The evidence shows that breast units have developed differing strategies for managing discordant reads and that the rationale for these is unclear. Consensus approaches encompass a diverse range of scenarios and it was not possible to establish the influence of component factors on decision-making.

In accordance with a scoping review's descriptive nature no attempt has been made to synthesise evidence on outcomes for these processes. The impact that the

arbitration or consensus process might have on breast unit's resources, the ability to maintain reporting standards, subsequent potential delays in patient pathways and cost effectiveness have not been explored.

No review of a comparable nature has been identified making this a novel study. Although the findings do not provide conclusive evidence on the effectiveness of different arbitration or consensus processes, they are valuable in providing a foundation upon which to build further knowledge (Pawson 2013).

5.8 Future Research or Recommendations

The scoping review has described the complexities within mammography reporting and that the UK is unique in that a consensus meeting may include a range of professionals (Radiologists, Advanced practitioners and Consultant radiographers). Concerns about the future availability of specialist breast radiologists have been highlighted (RCR 2016) and some services may already have or imminently have a single breast-screening radiologist. With the predicted retirement of the most experienced radiologists, units may soon have a workforce where advanced and consultant practitioners possess a substantial knowledge base relative to a newly appointed radiologist.

As the NHSBSP guidance on arbitration personnel is currently under review, it presents an opportune time to consider the qualities and characteristics required by an individual to undertake decision-making effectually. Under current NHSBSP

guidance a consultant radiographer can lead an assessment clinic directing additional imaging, performing appropriate interventions or ultimately discharging the lady to routine screening. The autonomous nature of the role extends to MDT meetings where advice can be given on patient management, but based on current clinical protocols arbitration of discrepant screen readings is considered a medic responsibility. Interestingly, a recent survey (Culpan 2016: 4) establishing the participation of UK radiographers in mammography image interpretation reports that 23% (15/66) of radiographers stated '*giving a third opinion or casting vote in discordant double reading*' of screening. This implies that units have already implemented changes in local practice through governance systems prior to the review of the guidance.

The primary aim of future research would be to establish current practice. For example, what do practitioners understand by the terms consensus and arbitration? and to develop clear precise definitions and guidance on the processes. A review of actual clinical practice could be ascertained via a national survey to establish a number of factors. What strategies are centres currently using for prevalent and incident screens, what professionals constitute consensus meetings, are specific times/days scheduled for arbitration or consensus, and which cases are reviewed (concordant and/or discordant only).

Further research would be required to:

1. Explore and explain how and why a system was established.

2. Explore the impact on the professionals of ad-hoc impromptu arbitration/consensus meetings and the effects on managing a service with such variable scheduling. Are individuals under pressure to make impromptu decisions as cases are about to breach targets? Is there an impact on the professional's performance relative to fatigue, time of day, immediate prior task?
3. Explore the clinical implications (time/resources/benefits) of a consensus panel reviewing all recall cases (concordant and discordant). Does the time and resources invested in reviewing concordant recalls result in significant reductions in recall rates?
4. Explore the dynamics of the professionals that constitute consensus meetings - Do the dynamics change relative to the individuals present i.e. grade and experience of the staff? Determine how the final decision is made – is this a majority decision, weighted by experience, or by profession?
5. Third person arbitration or consensus lead is traditionally a medic responsibility. Given the current NHSBSP arbitration guidance review, if delegation to non-medics is recommended, further research may be required to ascertain why some sites will be early adopters and some sites possibly non-adopters of the guidance. Would this relate to varying professional principles across organisations and more notably to determine the consequences of disparate practice not just for professionals, but service users? In particular, to determine the impact on outcomes of performance measures (recall rates, PPV, screen to routine recall and screen to

assessment). Would radiographer arbitration result in an increase in recall rates as seen in the NDROR trial, albeit it was considered insignificant? It will also be important to evaluate the impact on the individuals of undertaking this task. Although it is likely to be the Director of Breast Screening unit's responsibility to delegate to competent individuals, do radiographers consider they are confident/experienced enough to take on the responsibility? Would they only favour leading a consensus group rather than undertaking an independent 3rd read? Would any radiographers decline the delegation? Or would they embrace the change.

A distinctive approach to further research would be to prospectively explore the Collective Intelligence theory proposed by Wolf et al (2015) which removes the documented problems associated with group-think, and aims to deliver improved performance over a solitary experienced arbitrator. However, it is recognised that this method requires multiple readers and may be hindered by the current and future workforce shortages. Does the future of breast screening reporting lie in a revolution of CAD systems set with varying thresholds to perform the initial evaluation of the images with only discrepant cases then being resolved by the human reporter?

References

Arksey, H and O' Malley K. (2005) 'Scoping studies; towards a methodological framework'. *International Journal of social Research Methodology* 8 (1), 19-32

Armstrong, R., Hall, BJ., Doyle, J., and Waters, E. (2011) 'Cochrane Update Scoping the scope of a Cochrane review'. *J Public Health* 33 (1), 147-150

Aromataris, E., and Riitano, D. (2014) 'Constructing a search strategy and searching for evidence'. *Am J Nurs*. 114 (5), 49–56

Autier, P., Boniol, M., LaVecchia, C., Vatten, L., and Gavin A. (2010) 'Disparities in breast cancer mortality trends between 30 European countries: retrospective trend analysis of WHO mortality database'. *BMJ* 341, pc3620

Aveyard, H. (2009) Doing a literature review in health and social care 2nd edn. Buckingham: Open University Press

Bankier, A.A., Levine, D., Halpern, E.F., and Kressel H.Y. (2010) 'Consensus interpretation in imaging research: is there a better way?' *Radiology* 257, 14–17

Barlow, W.E., Chi, C, and Carney, P.A. (2004) 'Accuracy of screening mammography interpretation by characteristics of radiologists'. *J Natl Cancer Inst* 96, 1840-1850

Bennett, RL., Sellars, SJ., Blanks, RG., and Moss, SM. (2012) 'An observational study to evaluate the performance of units using two radiographers to read screening mammograms'. *Clinical Radiology* 67, 114-121

Berg, W.A., Campassi, C., Langenberg, P., and Sexton M.J. (2000) 'Breast Imaging Reporting and Data System: inter- and intraobserver variability in feature analysis and final assessment'. *AJR Am J Roentgenol* 174, 1769–1777.

Berg, W.A., D'Orsi, C.J., and Jackson, V.P. (2002) 'Does training in the Breast Imaging Reporting and Data System (BI-RADS) improve biopsy recommendations or feature analysis agreement with experienced breast imagers at mammography?' *Radiology* 224, 871–880.

Betran, A.P., Say, L., Glmezoglu, A.M., Allen, T., and Hampson L. (2005) 'Effectiveness of different databases in identifying studies for systematic reviews: experience from the WHO systematic review of maternal morbidity and mortality'. *BMC Medical Research Methodology* 5, 6.

Bettany-Saltikov, J. (2010) 'Learning how to undertake a systematic review: part 1'. *Nursing Standard* 24 (50), 47–55

Bettany-Saltikov, J. (2010) 'Learning how to undertake a systematic review: part 2'. *Nursing Standard* 24 (51), 47–56

Birdwell, R.L. (2009) 'The preponderance of evidence supports computer-aided detection for screening mammography'. *Radiology* 253 (1), 9-16

Blanks, R. G., Wallis, M. G. & Moss, S. M. (1998) 'A comparison of cancer detection rates achieved by breast cancer screening programmes by number of readers, for one and two view mammography: Results from the UK national health service breast screening programme'. *Journal of Medical Screening* 5, 195-201.

Bond, M., Garside, R., and Hyde C. (2015) 'Improving screening recall services for women with false-positive mammograms: a comparison of qualitative evidence with UK guidelines'. *BMJ Open* 2015, 5:e005855

Booth, A., Papaioannou, D., and Sutton, A: (2012) Systematic approaches to a successful literature review, Sage, London

Buist, D.S., Anderson, M.L., and Haneuse, S.J. (2011) 'Influence of annual interpretive volume on screening mammography performance in the United States'. *Radiology* 259, 72-84

Cates, C.J., Stovold, E., and Welsh, E.J. (2014) 'How to make sense of a Cochrane systematic review'. *Breathe* 10 (2), 134-144

Caumo, F., Brunelli, S., Tosi, E., Teggi, S., Bovo, C., Bonavina, G. and Ciatto, S. (2011) 'On the role of arbitration of discordant double readings of screening

mammography: experience from two Italian programmes'. *Radiologia Medica* 116, 84-91

Cawson, J. N., Nickson C., Amos, A., Hill, G., Whan, A. B. & Kavanagh, A. M. (2009) 'Invasive breast cancers detected by screening mammography: A detailed comparison of computer-aided detection-assisted single reading and double reading'. *Journal of Medical Imaging and Radiation Oncology* 53, 442-449

Ciatto, S., Ambrogetti, D., Risso, G., Catarzi, S., Morrone, D., Mantellini, P. and Rosselli Del Turco, M. (2005) 'The role of arbitration of discordant reports at double reading of screening mammograms'. *Journal of Medical Screening* 12, 125-127

Coad, J., Hardicre, J., and Devitt, P. (2006) 'Searching for and using grey literature'. *Nursing Times*, 102 (50), 35-36

Cornford, E. J., Evans, A. J., James, J. J., Burrell, H. C., Pinder, S. E. and Wilson, A. R. M. (2005) 'The pathological and radiological features of screen-detected breast cancers diagnosed following arbitration of discordant double reading opinions'. *Clinical Radiology* 60,1182-1187

Crossan, M.M., and Apaydin, M. (2010) 'A multi-dimensional framework of organizational innovation: A systematic review of the literature'. *Journal of Management Studies* 47 (6), 1154–1191

Culpan, A.M. (2016) 'Radiographer involvement in mammography image interpretation: A survey of United Kingdom practice' *Radiography article in press*, 1-7

Cummings, G.C., MacGregor, T., Davey, M., Lee, H., Wong, C.A., Lo, E., Muise, M., and Stafford, E. (2010) 'Leadership styles and outcome patterns for the nursing workforce and work environment: A systematic review'. *International Journal of Nursing Studies* 47, 363–385

Debono, JC., Poulos, AE., Houssami, N., Turner, RM. and Boyages, J. (2015) 'Evaluation of radiographers' mammography screen-reading accuracy in Australia'. *Journal of Medical Radiation Sciences* 62, 15-22

Department of Health (DH) (2013) Public health functions to be exercised by NHS England Service Specification No.24 Breast Screening Programme

Department of Health (DH) (2011). Guide to waiting times (online). Available from: <http://nhs.uk/choiceintheNHS/Rightsandpledges/Watingtimes/Pages/Guide%20to%20waiting %20times.aspx > [28 Oct 2015]

Department of Health (2008) About the Department of Health. London available from: http://www.dh.gov.uk/en/Aboutus/index.htm [15 March 2016]

Department of Health (DH) (2007a). New Ways of Working for Everyone: A best practice guide. London: DH.

Department of Health (2000). 'The NHS cancer plan. A plan for investment. A plan for reform'. London: DH

Derry, S., Loke, Y.K., Aronson, J.K. (2001). Incomplete evidence: the inadequacy of databases in tracing published adverse drug reactions in clinical trials. *BMC Med Res* Methodol 1.

Dinnes, J., Moss, S., Melia, J., Blanks, R., Song, F. and Kleijnen, J. (2001) 'Effectiveness and cost-effectiveness of double reading of mammograms in breast cancer screening: Findings of a systematic review'. *Breast* 10, 455-463

Duijm, L.E., Louwman, M.W., Groenewoud, J.H., van de Poll-Franse, L.V., Fracheboud, J, and Coebergh, J.W. (2009) 'Interobserver variability in mammography screening and effect of type and number of readers on screening outcome'. *Br J Cancer* 100, 901-907

Duijm, L. E. M., Groenewoud, J. H., Hendriks, J. H. C. L. and De Koning, H. J. (2004) 'Independent Double Reading of Screening Mammograms in the Netherlands: Effect of Arbitration Following Reader Disagreements'. *Radiology* 231, 564-570.

Ferlay, J., Soerjomataram, I., Ervik, M., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, DM., Forman, D. and Bray, F. (2012) 'Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11' [Internet]. Lyon, France: International Agency for Research on Cancer; 2013. Available from < http://globocan.iarc.fr,> [10 May 2016]

Forest, A.P.M. (1986) 'Breast Cancer Screening: report to the health ministers of England, Wales, Scotland and Northern Ireland. London: HMSO

Gilbert, F.J., Tucker, L., Gillan, M.G., Willsher, P., Cooke, J., Duncan, K.A., Michell, M.J., Dobson, H.M., Lim, Y.Y., Purushothaman, H., Strudley, C., Astley, S.M., Morrish, O., Young, K.C., and Duffy, S.W. (2015) 'The TOMMY trial: a comparison of TOMosynthesis with digital MammographY in the UK NHS Breast Screening Programme--a multicentre retrospective reading study comparing the diagnostic performance of digital breast tomosynthesis and digital mammography with digital mammography alone'. *Health Technol Assess*. 19 (4), 1-136

Godin, K., Stapleton, J., Kirkpatrick, S.I., Hanning, R.M., and Leatherdale, S.T. (2015) 'Applying systematic review search methods to the grey literature: a case study examining guidelines for school-based breakfast programs in Canada' *BioMed Central* 4, 138

Goldacre, B (2012) Bad Pharma: How drug companies mislead doctors and harm patients. London: Fourth estate.

Greenhalgh, T and Peacock, R. (2005) 'Effectiveness and efficiency of search methods in systematic reviews of complex evidence: audit of primary sources.' *British Medical Journal* 331, 1064-1065

Groenewoud, J. H., Otten, J. D. M., Fracheboud, J., Draisma, G., Van Ineveld, B. M., Holland, R., Verbeek, A. L. M. and De Koning, H. J. (2007) 'Cost-effectiveness of different reading and referral strategies in mammography screening in the Netherlands'. *Breast Cancer Research and Treatment* 102, 211-218

Hemingway, P. and Brereton, N. (2009) What is a Systematic Review. What is…? series, 2nd edition. Hayward Medical Communications, Hayward Group Ltd.

Hodgson, R., Heywang-Kobrunner, S.H., Harvey, S.C., Edwards, M., Shaikh, J., Arber, M and Glanville, J. (2016) 'Systematic review of 3D mammography for breast cancer screening' *The Breast* 27, 52-61

Hofvind, S., Geller, B. M., Rosenberg, R. D. and Skaane, P. (2009) 'Screening-detected Breast Cancers: Discordant Independent Double Reading in a Population-based Screening Program'. *Radiology* 253, 652-660.

Hukkinen, K., Kivisaari, L. and Vehmas, T. (2006) 'Impact of the number of readers on mammography interpretation'. *Acta Radiologica* 47, 655-659.

James, J. J. and Cornford E. J. (2009) ‘Does computer-aided detection have a role in the arbitration of discordant double-reading opinions in a breast-screening programme?’ *Clinical Radiology*, 64, 46-51

Jenkins, J., Murphy, A. E., Edmondson-Jones, M., Sibbering, D. M. and Turnbull, A. E. (2014) ‘Film reading in the East Midlands Breast Screening Programme - Are we missing opportunities for earlier diagnosis?’ *Clinical Radiology* 69, 385-390

Kable, A., Pich, J. and Maslin-Prothero, S. (2012) 'A structured approach to documenting a search strategy for publication: A 12 step guideline for authors'. *Nurse Education Today* 32, 878-886

Kerr, N.L., and Tindale, R.S. (2004) 'Group performance and decision making’ *Annual Review of Psychology* 55, 623–655

Khoo, L. A. L., Taylor, P. and Given-Wilson, R. M. (2005) 'Computer-aided detection in the United Kingdom National Breast Screening Programme: Prospective study'. *Radiology* 237, 444-449

Klompenhouwer, E. G., Voogd, A. C., Den Heeten, G. J., Strobbe, L. J. A., Tjan-Heijnen, V. C., Broeders, M. J. M. and Duijm, L. E. M. (2015a) 'Discrepant screening mammography assessments at blinded and non-blinded double reading: impact of arbitration by a third reader on screening outcome'. *European Radiology* 25, 2821-2829.

Klompenhouwer, E. G., Weber, R. J. P., Voogd, A. C., Den Heeten, G. J., Strobbe, L. J. A., Broeders, M. J. M., Tjan-Heijnen, V. C. G. and Duijm, L. E. M. (2015b) 'Arbitration of discrepant BI-RADS 0 recalls by a third reader at screening mammography lowers recall rate but not the cancer detection rate and sensitivity at blinded and non-blinded double reading'. *Breast* 24, 601-607

Labin, S., Duffy, J., Meyers, D., Wandersman, A. and Lesesne, C. (2012) 'A Research Synthesis of the Evaluation Capacity Building Literature'. *American Journal of Evaluation* 33, 307-338

Laming, D., and Warren, R. (2000) 'Improving the detection of cancer in the screening of mammograms'. *J Med Screen* 7, 24–30

Lang, K., Andersson, I., Rosso, A., Tingberg, A., Timberg, P. and Zackrisson, S. (2016) 'Performance of one-view breast tomosynthesis as a stand-alone breast cancer screening modality: results from the Malmo Breast Tomosynthesis Screening Trial, a population-based study'. *European Radiology* 26, 184-190

Lehamn, C.D., Wellman, R.D., Buist, D.S., Kerlikowske, K., Tosteson, A.N. and Migloretti, D.L. (2015) 'Diagnostic Accuracy of Digital Screening Mammography With and Without Computer-Aided Detection'. *JAMA Intern Med* 175, (11), 1828-1837

Levac, D., Colquhoun, H., O'Brien, K.K. (2010). 'Scoping studies: advancing the methodology'. *Implement Sci* 5, 69.

Liston, J.C., and Dall B.J. (2003) 'Can the NHS Breast Screening Programme afford not to double read screening mammograms?' *Clin Radiol* 58 (6) 474-477

Lloyd-Jones, N., and Masterton, A., (2010) Writing skills and developing an argument, In: Maslin-Prothero, S. (Ed.), Bailliere's study skills for nurses and midwives, 4th ed. Edinburgh, Bailliere Tindall.

Local cancer intelligence (2016) [online] <http://lci.cancertoolkit.co.uk/> [12 April 2016]

Mahmood, Q., Eerd, D.V. and Irvin, E. (2014) 'Searching for grey literature for systematic reviews: challenges and benefits. *Res Synthesis Methods* 5(3), 221–34

Marmot, M., Altman, D., Cameron, D., Dewar, J., Thompson, S., Wilcox, M. and The Independent UK Panel on Breast Cancer Screening. (2013) 'The benefits and harms of breast cancer screening: an independent review'. *British Journal of Cancer*, 108, 2205-2240.

Maslin-Prothero, S., and Bennion, A. (2010) 'Integrated team working: a literature review'. *International Journal of Integrated Care* 10, 1–11.

Matcham, N.J., Ridley, N.T., Taylor, S.J., Cook, J.L., Scolding, J. (2004) Breast screening: the use of consensus opinion for all recalls' *Breast* 13 (3), 184-187

McDermott, O., Crellin, N., Ridder, H. M., and Orrell, M. (2013) 'Music therapy in dementia: a narrative synthesis systematic review'. *International journal of geriatric psychiatry* 28 (8), 781-794.

McGowan, J., and Sampson, M., (2005). 'Systematic reviews need systematic searchers'. *Journal of the Medical Library Association* 93 (1), 74–80

Miglioretti, D.L., Gard, C.C., and Carney, P.A. (2009) 'When Radiologists Perform Best: The Learning Curve in Screening Mammogram Interpretation'. *Radiology* 253, 632-640

Moran, S., and Warren-Forward, H. (2016) 'The diagnostic accuracy of radiographers assessing screening mammograms: A systematic review'. *Radiography* 22, 137-146

Moser, K., Sellars, S., Wheaton, M., Cooke, J., Duncan, A., Maxwell, A., Michell, M., Wilson, M., Beral, V., Peto, R., Richards, M., and Patnick, J. (2011) 'Extending the age range for breast screening in England: pilot study to assess the feasibility and acceptability of randomization'. *J Med Screen* 18 (2), 96-102.

Mucci, B., Athey, G. and Scarisbrick, G. (1999) 'Double read screening mammograms: The use of a third reader to arbitrate on disagreements'. *Breast* 8, 37-39

Mushlin, A.I., Kouides, R.W., and Shapiro, D.E. (1998) 'Estimating the accuracy of screening mammography: a meta-analysis'. *Am J Prev Med*. 14 (2), 143-53.

National Academies of Sciences, Engineering, and Medicine. 2015. Assessing and improving the interpretation of breast images: Workshop Summary. Washington, DC: The National Academies Press.

NCIN. 2012. Cancer Commissioning Toolkit [online]. Available from https://www.cancertoolkit.co.uk/Charts/Expenditure/SingleYearCancerShareSpendBySHA [20 November 2015]

NHS Cancer Screening Programmes Advisory Committee. (2011) 'Quality Assurance Guidelines for Breast Screening Radiology'. Publication 59

NHS Cancer Screening Programmes Advisory Committee. (2010) 'Clinical guidelines for healthcare professionals screening women for breast cancer'. Publication 49

Office for National Statistics, (2015). Cancer Registration Statistics, England, 2013.

Pauli, R., Hammond, S., Cooke, J. and Ansell, J. (1996) 'Comparison of radiographer/radiologist double film reading with single reading in breast cancer screening'. J Med Screen 3, 18-22.

Perry, N., Broeders, M., de Wolf, C., Tornberg, S., Holland, R., and von Karsa, L. (2008) 'European guidelines for quality assurance in breast cancer screening and diagnosis'. Fourth edition–summary document. *Ann Oncol*, 19 (4), 614–22.

Popay, J., Roberts, H., Sowden, A., Petticrew, M., Arai, L., Rodgers, M., and Duffy, S. (2006). 'Guideline on the conduct of narrative synthesis in systematic reviews'. A product from the ESRC methods programme. Version, 1.

Posso, M. C., Puig, T., Quintana, M. J., Sola-Roca, J. and Bonfill, X. (2016) 'Double versus single reading of mammograms in a breast cancer screening programme: a cost-consequence analysis'. *European Radiology* 1-10.

Royal College of Radiologists (2016) The breast imaging and diagnostic workforce in the United Kingdom. Results of a survey of NHS Breast Screening Programme units and radiology departments. BFCR(16)2

Shaw, C. M., Flanagan, F. L., Fenlon, H. M. and Mcnicholas, M. M. (2009) 'Consensus Review of Discordant Findings Maximizes Cancer Detection Rate in Double-Reader Screening Mammography: Irish National Breast Screening Program Experience'. *Radiology* 250, 354-362

Skaane, P., Bandos, A. I., Gullien, R., Eben, E. B., Ekseth, U., Haakenaasen, U., Izadi, M., Jebsen, I. N., Jahr, G., Krager, M. and Hofvind, S. (2013a) 'Prospective trial comparing full-field digital mammography (FFDM) versus combined FFDM and

tomosynthesis in a population-based screening programme using independent double reading with arbitration'. *European Radiology* 23, 2061-2071

Skaane, P., Bandos, A. I., Gullien, R., Eben, E. B., Ekseth, U., Haakenaasen, U., Izadi, M., Jebsen, I. N., Jahr, G., Krager, M., Niklason, L. T., Hofvind, S. and Gur, D. (2013b) 'Comparison of digital mammography alone and digital mammography plus tomosynthesis in a population based screening program'. *Radiology* 267, 47-56

Skaane, P., Diekmann, F., and Balleyguier, C. (2008) 'Observer variability in screen-film mammography versus full field digital mammography with soft-copy reading'. *Eur Radiol* 18, 1134-1143

Skaane, P., Hofvind, S. and Skjennald, A. (2007) 'Randomized trial of screen-film versus full-field digital mammography with soft-copy reading in population-based screening program: Follow-up and final results of Oslo II study'. *Radiology* 244, 708-717

Smith-Bindman, R., Chu, P.W., Miglioretti, D.L., Sickles, E.A., Blanks, R., Ballard-Barbash, R., Bobo, J.K., Lee, N.C., Wallis, M.G., Patnick, P., and Kerlikowske, K. (2003) 'Comparison of screening mammography in the United States and the United Kingdom'. *JAMA* 290, 2129-2137

Taylor, K., Britton, P., O'Keeffe, S. and Wallis, M. (2011) 'Quantification of the UK 5-point breast imaging classification and mapping to BI-RADS to facilitate comparison with international literature'. *British Journal of Radiology* 84, 1005-1010

Taylor-Philips S et al (2011) 'The time course of cancer detection performance' Loughborough University Institutional Repository. Available from: < https://dspace.lboro.ac.uk/2134/8250 > accessed [10 May 2016]

Taylor, P. and Potts, H. W. W. (2008) 'Computer aids and human second reading as interventions in screening mammography: Two systematic reviews to compare effects on cancer detection and recall rate'. *European Journal of Cancer* 44, 798-807

The Joanna Briggs Institute Reviewers' Manual. (2015) Methodology for JBI Scoping Reviews Joanna Briggs Institute.

Theberge, I., Chang, SL., and Vandal, N. (2014) 'Radiologist interpretive volume and breast cancer screening accuracy in a Canadian organized screening program'. *J Natl Cancer Inst* 106, 461.

Timmins, F., and McCabe, C. (2005). 'How to conduct an effective literature search'. *Nursing Standard* 20 (11), 41–47

Torres-Mejia, G., Smith, R. A., Carranza-Flores, L., Bogart, A., Martinez-Matsushita, L., Migloretti, D. L., Kerlikowske, K., Ortega-Olvera, C., Montemayor-Varela, E., Angeles-Llerenas, A., Bautista-Arredondo, S., Sanchez-Gonzalez, G., Martinez-Montanez, O. G., Uscanga-Sanchez, S. R., Lazxano-Ponce, E. and Hernandez-Avila, M. (2015). 'Radiographers supporting radiologists in the interpretation of screening

mammography: a viable strategy to meet the shortage in the number of radiologists'. *BMC Cancer* 15, 410.

Wai, C. J., Al-Mubarak, G., Homer, M.J., Goldkamp, A., Samenfeld-Specht, M., Lee, Y., Logvinenko, T., Rothschild, J.G. and Graham, R.A. (2013) 'A modified triple test for palpable breast masses: the value of ultrasound and core needle biopsy'. *Annals of surgical oncology* 20 (3), 850-855

Whiting, P., Savovic, J., Higgins, J. P., Caldwell, D.M., Reeves, B.C., Shea, B., Davies, P., Kleijnen, J., Churchill, R., (2016c) 'ROBIS: A new tool to assess risk of bias in systematic reviews was developed'. *J Clin Epidemiol* 69, 225-234

Whiting, P., Harbord, R. and Kleijnen, J. (2005b) 'No role for quality scores in systematic reviews of diagnostic accuracy studies'. *BMC Med Res Methodol* 5, 19

Whiting, P., Rutjes, A.W.S., Reitsma, J.B., Bossuyt, P.M.M. and Kleijnen, J. (2003a) 'The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews'. *BMC Med Res Methodol* 3, 25

Wolf, M., Krause, J., Carney, P. A., Bogart, A. and Kurvers, R. H. J. M. (2015) 'Collective intelligence meets medical decision-making: The collective outperforms the best radiologist'. *PLoS ONE*, 10

Appendix A. NHSBSP Quality Standards

Objective	Criteria	Minimum standard	Achievable standard
1. To maximise the number of eligible women who attend for screening	The percentage of eligible women who attend for screening	>70% of invited women to attend for screening	80%
2. To maximise the number of cancers detected	a) The rate of invasive cancers detected in eligible women invited and screened b) The rate of cancers detected that are in situ carcinoma c) Standardised detection ratio (SDR)	Prevalent screen >3.6 per 1,000 Incident screen >4.1 per 1,000 Prevalent screen >0.5 per 1,000 Incident screen >0.6 per 1,000 > 1.0	Prevalent screen >5.1 per 1,000 Incident screen >5.7 per 1,000 >1.4
3. To maximise the number of small invasive cancers detected	The rate of invasive cancers less than 15 mm in diameter detected in eligible women invited and screened	Prevalent screen >2.0 per 1,000 Incident screen > 2.3 per 1,000	Prevalent screen >2.8 per 1,000 Incident screen >3.1 per 1,000
4. To achieve optimum image quality	a) High contrast spatial resolution b) Minimal detectable contrast 5-6 mm detail 0.5 mm detail 0.25 mm detail c) Optical density	>12 lp/mm < 1.2% < 5% <8%	 <0.8% < 3% <5%
5. To limit radiation dose	Mean glandular dose per exposure for a standard breast at clinical settings	<2.5 mGy	
6. To minimise the number of women undergoing repeat examinations	The number of repeat examinations	<3% of total examinations	<2% of total examinations
7. To minimise the number of women screened who are referred for further tests	a) The percentage of women who are referred for assessment b) The percentage of women screened who are placed on short term recall	Prevalent screen <10% Incident screen <7% <0.25%	Prevalent screen <7% Incident screen <5%

8. To ensure that the majority of cancers, both palpable and impalpable, receive a non-operative tissue diagnosis of cancer	(a) The percentage of women who have a non-operative diagnosis of cancer by cytology or needle histology after a maximum of two visits (b) The percentage of women who have a non-operative diagnosis of DCIS by cytology or needle histology after a maximum of two attempts	>90% >85%	>95% >90%
9. To minimise the number of unnecessary operative procedures	The rate of benign biopsies	Prevalent screen <1.5 per 1,000 Incident screen <1.0 per 1,000	Prevalent screen <1.0 per 1,000 Incident screen <0.75 per 1,000
10.To minimise the number of cancers presenting between screening episodes in the women screened	The rate of cancers presenting in screened women a) in the two years following a normal screening episode b) in the third year following a normal screening episode	Expected standard 1.2 per 1,000 women screened in the first two years 1.4 per 1,000 women screened in the third year	
11.To ensure that women are recalled for screening at appropriate intervals	The percentage of eligible women whose first offered appointment is within 36 months of their previous screen	>90%	100%
12.To minimise anxiety for women who are awaiting the results of screening	The percentage of women who are sent their result within two weeks	>90%	100%
13.To minimise the interval from the screening mammogram to assessment	The percentage of women who attend an assessment centre within three weeks of attendance for the screening mammogram	>90%	100%
14.To minimise diagnostic delay for women who are diagnosed non-operatively	Proportion of women for whom the time interval between non-operative biopsy and result is one week or less	>90%	100%

15.To minimise the delay for women who require surgical assessment	Proportion of women for whom the time interval between the decision to refer to a surgeon and surgical assessment is one week or less	>90%	100%
16.To minimise the delay between referral for investigation and first breast cancer treatment	The percentage of women who are admitted for treatment within two months of the date of referral	>90%	100%

Appendix B. Initial Keywords and Subject Headings

Criteria	Where to search	Keywords
>5000 films per year including 1500 first reads, 4000 screening mammograms	Primary database's CENTRAL CINAHL DARE EMBASE MEDLINE NHS EED NICE NHS HMIC PsycINFO Pubmed Cochrane Database of Systematic Reviews. Web of science Popline Grey literature will also be searched: conference proceedings, theses dissertations', book citations, websites of NHS England, NHS Breast Cancer Screening Programme; Professional body publications/guidance, HEE, NCIN, NHSBSP cancer screening. Searches will be limited to English language only.	Accuracy Attention Breast Neoplasms/*radiography/*diagnostic/*pathology *Clinical Competence Clinical Protocols Diagnostic errors/*prevention & control/*trends Diagnostic errors/*prevention & control/stats and numerical data Diagnostic imaging Double-Blind Method Early detection of cancers/*methods False negative reactions False positive reactions Fatigue Female Film readers Film reading volumes Humans Image interpretation *Mammography*/*standards Mass Screening/*methods/*standards National Health programs/classification Observer Variation Observer performance Optimal Pattern Recognition, Visual Perception Predictive Value of Tests Prognosis Radiographer *Radiology Radiography/standards Reproducibility of Results *Research Design Retrospective studies Screen-reading Sensitivity and specificity Task Performance and Analysis Volume Vigilance decrement
> 2yrs film reading experience in breast screening. New Radiologist - Full appropriate training, ideally undertaken a breast fellowship		Breast Clinical Competence/standards Clinical Protocols Experience Fellowship Film readers Film reading volumes Humans *Mammography Mass Screening/*methods Neoplasms/*radiography Observer performance Observer Variation Reproducibility of Results Task Performance and Analysis Training*
Full participation in assessment clinics including decision making		Attention Breast Neoplasms/*radiography/*diagnosis Breast triple assessment* *Clinical Competence

		Clinical practice guidelines Clinical Protocols Clinical strategies Clinical decision analysis Critical thinking Decision aids Decision making* Diagnostic Errors/*prevention & control/*trends Fatigue Humans Mammography/*methods Mammography/psychology Pattern Recognition, Visual Task Performance and Analysis Threshold Vigilance decrement
Regular attender/participant at MDT. Desirable > 20 per year		Breast Neoplasms/*radiography Clinical Protocols Health care team Hospital /hospital setting/hospital based Integrated care Interdisciplinary/team Medical care team Multidisciplinary Multidisciplinary team Peer review measures Patient care team Person-centred care, Secondary care
Regular audit personal and team results, Reflective learning		Clinical Audit *Clinical Competence Clinical Protocols Continuing education Distance learning Education, Distance Education, Medical, Continuing Educational Measurement Evaluation studies Mammography Personal audit Questionnaires Quality improvement Reducing recall rates Reflection Reflective learning Radiology/education Task Performance and Analysis
Review of interval cancers, screen detected cancers Participation in PERFORMS		Accuracy Attention Breast Neoplasms/*diagnosis Cancer detection rates *Clinical Competence Diagnostic Errors/*prevention & control/*statistics & numerical data Diagnostic Imaging/*statistics & numerical data Diagnostic Errors/*prevention & control/*trends False Negative Reactions Fatigue Humans Interval cancer Mammography/*methods Mass Screening MEDLINE/*statistics & numerical data Missed cancers Neoplasms/*radiography/*diagnosis/*pathology Pattern Recognition, Visual PERFORMS

		Performance Population screening Predictive Value of Tests Prognosis Program evaluation Quality indicators, Health care* Quality Improvement/*statistics & numerical data Radiology/*education/*statistics & numerical data Radiological review Reproducibility of Results Retrospective Studies Task Performance and Analysis Time Factors Treatment Outcome Vigilance decrement
Ongoing professional development and appraisal, SCoR 4 pillars of consultant radiographer practice		Advanced Practice/Radiography* Allied Health Personnel/education Appraisal *Clinical Competence Clinical Protocols Continued Professional Development Continuing Medical Education Education, Distance/organisation & administration Education medical Health professional education Health services research/methods Humans Practicing health professionals Health professional students Mammography Models, Educational Universities/organization & administration Organisational innovation Physician–radiographer relations Practice guidelines as Topic/standards* Professional delegation *Radiographer Retention SCoR Skills development Systematic literature review

Appendix C. Pubmed Search

Recent queries in Pubmed		
Search	**Query**	**Items found**
#1	Search ("breast neoplasm" or "breast carcinoma" or "breast tumour' or "breast tumor" or "breast cancer")	8796
#2	Search ("mass screening" or "breast scan"or "breast screen" or " breast radiograph" or "breast imaging" or "breast visualise" or "breast test" or "breast mammogram" or "breast diagnosis")	90602
#3	Search ("mammogram" or "mammography" or)	31899
#4	Search (#2 or #3)	113648
#5	Search ("early detection of cancer" or "National health service breast screening program" or "NHSBSP" or "UK breast screening program" or "NHS breast screening program")	12665
#6	Search (#1 and #4 and #5)	12
#7	Search ("arbitration" or "negotiation" or "discordance" or "discrepancy" or "disparity" or "disagree" or "conflict" or "different" or " inconsistent" or "variation" or " consensus" or "uncertain")	2994590
#8	Search (#6 and #7)	5

Appendix D. Medline Arbitration Search

1. Medline; exp BREAST NEOPLASMS/; 233023 results
2. Medline; (breast adj3 (neoplasm* OR carcinoma* OR tumour* OR tumor* OR cancer*)).ti,ab; 238107results
3. Medline; 1 OR 2; 296042 results
4. Medline; exp MASS SCREENING/; 105328 results
5. Medline; (breast adj3 (scan* OR screen* OR radiograph* OR imaging OR visualise OR visualize OR
exam* OR test* OR mammogra* OR routine* OR check* OR diagnos* OR detect*)).ti,ab; 49354 results
6. Medline; (mammogra* adj3 (scan* OR screen* OR visualise OR visualize OR exam* OR test* OR
breast*)).ti,ab; 14103 results
7. Medline; exp MAMMOGRAPHY/; 25226 results
8. Medline; 4 OR 5 OR 6 OR 7; 159997 results
9. Medline; "Early detection of cancer*".ti,ab; 953 results
10. Medline; ("National Health Service Breast Screening Program" OR "NHSBSP").ti,ab; 86 results
11. Medline; "UK breast screen* program*".ti,ab; 35 results
12. Medline; "NHS breast screen* program*".ti,ab; 107 results
13. Medline; 9 OR 10 OR 11 OR 12; 1150 results
14. Medline; 3 AND 8 AND 13; 287 results
15. Medline; exp NEGOTIATING/; 5256 results
16. Medline; (arbitration* OR arbitrat* OR discordan* OR discrepan* OR disparity* OR negotiat* OR
disagree* OR conflict* OR differen* OR inconsisten* OR variation* OR consensus* OR uncertain*).ti,ab;
5001928 results
17. Medline; 15 OR 16; 5004649 results
18. Medline; 14 AND 17; 96 results
19. Medline; 18 [Limit to: (Language English) and Humans]; 84 results

Appendix E. Medline Decision Making Search

Search History:
1. Medline; exp BREAST NEOPLASMS/; 234067 results.
2. Medline; (breast adj3 (neoplasm* OR carcinoma* OR tumour* OR tumor* OR cancer*)).ti,ab; 38575 results.
3. Medline; 1 OR 2; 242222 results.
4. Medline; exp MASS SCREENING/; 105729 results.
5. Medline; (breast adj3 (scan* OR screen* OR radiograph* OR imaging OR visualise OR visualize OR exam* OR test* OR mammogra* OR routine* OR check* OR diagnos* OR detect*)).ti,ab; 10840 results.
6. Medline; (mammogra* adj3 (scan* OR screen* OR visualise OR visualize OR exam* OR test* OR breast*)).ti,ab; 8240 results.
7. Medline; exp MAMMOGRAPHY/; 25306 results.
8. Medline; 4 OR 5 OR 6 OR 7; 129630 results.
9. Medline; "Early detection of cancer*".ti,ab; 957 results.
10. Medline; ("National Health Service Breast Screening Program" OR "NHSBSP").ti,ab; 86 results.
11. Medline; "UK breast screen* program*".ti,ab; 0 results.
12. Medline; "NHS breast screen* program*".ti,ab; 0 results.
13. Medline; 9 OR 10 OR 11 OR 12; 1043 results.
14. Medline; 3 AND 8 AND 13; 135 results.
15. Medline; exp DECISION MAKING/; 153263 results.
16. Medline; ("decision mak* OR shared decision making" OR "medical decision making" OR "choice behaviour" OR "problem solving" OR "clinical decision analysis" OR "critical think*" OR "decision aids" OR "Task performance and analysis").ti,ab; 19124 results.
17. Medline; 15 OR 16; 170017 results.
18. Medline; 14 AND 17; 2 results.
19. Medline; 18 [Limit to: (Language English) and Humans]; 2 results.

Appendix F. EMBASE Arbitration Search

Search History:
1. EMBASE; exp BREAST CANCER/; 327328 results.
2. EMBASE; exp BREAST TUMOR/; 395504 results.
3. EMBASE; (breast adj3 (neoplasm* OR carcinoma* OR tumour* OR tumor* OR cancer*)).ti,ab; 320225 results.
4. EMBASE; 1 OR 2 OR 3; 437065 results.
5. EMBASE; exp MASS SCREENING/; 176405 results.
6. EMBASE; (breast adj3 (scan* OR screen* OR radiograph* OR imaging OR visualise OR visualize OR exam* OR test* OR mammogra* OR routine* OR check* OR diagnos* OR detect*)).ti,ab; 55378 results.
7. EMBASE; (mammogra* adj3 (scan* OR screen* OR visualise OR visualize OR exam* OR test* OR breast*)).ti,ab; 15742 results.
8. EMBASE; exp DIGITAL MAMMOGRAPHY/ OR exp MAMMOGRAPHY/; 44056 results.
9. EMBASE; 5 OR 6 OR 7 OR 8; 241733 results.
10. EMBASE; "Early detection of cancer*".ti,ab; 1295 results.
11. EMBASE; ("National Health Service Breast Screening Program" OR "NHSBSP").ti,ab; 159 results.
12. EMBASE; "UK breast screen* program*".ti,ab; 51 results.
13. EMBASE; "NHS breast screen* program*".ti,ab; 158 results.
14. EMBASE; 10 OR 11 OR 12 OR 13; 1617 results.
15. EMBASE; 4 AND 9 AND 14; 443 results.
16. EMBASE; exp INTERPERSONAL COMMUNICATION/; 450801 results.
17. EMBASE; (arbitration* OR arbitrat* OR discordan* OR discrepan* OR disparity* OR negotiat* OR disagree* OR conflict* OR differen* OR inconsisten* OR variation* OR consensus* OR uncertain*).ti,ab; 6118002 results.
18. EMBASE; 16 OR 17; 6434718 results.
19. EMBASE; 15 AND 18; 162 results.
20. EMBASE; 19 [Limit to: Human and (Languages English)]; 142 results.
21. EMBASE; 20 [Limit to: Human and (Languages English) and Publication Year 2005-2016]; 101 results.

Appendix G. EMBASE Decision Making Search

Search history:
1. EMBASE; exp BREAST CANCER/; 325856 results
2. EMBASE; exp BREAST TUMOR/; 393803 results
3. EMBASE; (breast adj3 (neoplasm* OR carcinoma* OR tumour* OR tumor* OR cancer*)).ti,ab; 318694 results
4. EMBASE; 1 OR 2 OR 3; 435037 results
5. EMBASE; exp MASS SCREENING/; 175689 results
6. EMBASE; (breast adj3 (scan* OR screen* OR radiograph* OR imaging OR visualise OR visualize OR exam* OR test* OR mammogra* OR routine* OR check* OR diagnos* OR detect*)).ti,ab; 55141 results
7. EMBASE; (mammogra* adj3 (scan* OR screen* OR visualise OR visualize OR exam* OR test* OR breast*)).ti,ab; 15697 results
8. EMBASE; exp DIGITAL MAMMOGRAPHY/ OR exp MAMMOGRAPHY/; 43936 results
9. EMBASE; 5 OR 6 OR 7 OR 8; 240771 results
10. EMBASE; "Early detection of cancer*".ti,ab; 1290 results
11. EMBASE; ("National Health Service Breast Screening Program" OR "NHSBSP").ti,ab; 158 results
12. EMBASE; "UK breast screen* program*".ti,ab; 49 results
13. EMBASE; "NHS breast screen* program*".ti,ab; 158 results
14. EMBASE; 10 OR 11 OR 12 OR 13; 1609 results
15. EMBASE; 4 AND 9 AND 14; 443 results
16. EMBASE; exp DECISION MAKING/; 241331 results
17. EMBASE; ("decision mak* OR shared decision making" OR "medical decision making" OR "choice behaviour" OR "problem solving" OR "clinical decision analysis" OR "critical think*" OR "decision aids" OR "Task performance and analysis").ti,ab; 23459 results
18. EMBASE; 16 OR 17; 259939 results
19. EMBASE; 15 AND 18; 7 results
20. EMBASE; 19 [Limit to: Human and English Language]; 7 results

Appendix H. CINAHL Search

Search History

1. CINAHL; exp BREAST NEOPLASMS/ OR exp CARCINOMA, LOBULAR/; 38776 results.
2. CINAHL; (breast adj3 (neoplasm* OR carcinoma* OR tumour* OR tumor* OR cancer*)).ti,ab; 29790 results.
3. CINAHL; 1 OR 2; 43220 results.
4. CINAHL; exp HEALTH SCREENING/; 46351 results.
5. CINAHL; (breast adj3 (scan* OR screen* OR radiograph* OR imaging OR visualise OR visualize OR exam* OR test* OR mammogra* OR routine* OR check* OR diagnos* OR detect*)).ti,ab; 8223 results.
6. CINAHL; (mammogra* adj3 (scan* OR screen* OR visualise OR visualize OR exam* OR test* OR breast*)).ti,ab; 2778 results.
7. CINAHL; exp MAMMOGRAPHY/ OR exp RESCREENING/ OR exp BREAST/; 8128 results.
8. CINAHL; 4 OR 5 OR 6 OR 7; 56692 results.
9. CINAHL; "Early detection of cancer*".ti,ab; 121 results.
10. CINAHL; ("National Health Service Breast Screening Program" OR "NHSBSP").ti,ab; 15 results.
11. CINAHL; "UK breast screen* program*".ti,ab; 8 results.
12. CINAHL; "NHS breast screen* program*".ti,ab; 44 results.
13. CINAHL; 9 OR 10 OR 11 OR 12; 182 results.
14. CINAHL; 3 AND 8 AND 13; 81 results.
15. CINAHL; exp ARBITRATION/; 252 results.
16. CINAHL; (arbitration* OR arbitrat* OR discordan* OR discrepan* OR disparity* OR negotiat* OR disagree* OR conflict* OR differen* OR inconsisten* OR variation* OR consensus* OR uncertain*).ti,ab; 370113 results.
17. CINAHL; 15 OR 16; 370244 results.
18. CINAHL; 14 AND 17; 15 results.
19. CINAHL; 18 [Limit to: (Language English)]; 14 results.

Appendix I. Cochrane Decision Making Search

Last Saved: 19/01/2016 16:14:17.733
Description:

ID	Search
#1	MeSH descriptor: [Breast Neoplasms] explode all trees
#2	breast adj3 (neoplasm* or carcinoma* or tumour* or tumor* or cancer*)
#3	#1 or #2
#4	MeSH descriptor: [Mass Screening] explode all trees
#5	breast adj3 (scan* or screen* or radiograph* or imaging or visualise or visualize or exam* or test* or mammogra* or routine* or check* or diagnos* or detect*)
#6	mammogra* adj3 (scan* or screen* or visualise or visualize or exam* or test* or breast*)
#7	MeSH descriptor: [Mammography] explode all trees
#8	#4 or #5 or #6 or #7
#9	"Early detection of cancer*"
#10	"National Health Service Breast Screening Program" or "NHSBSP"
#11	"UK breast screen* program*"
#12	"NHS breast screen* program*"
#13	#9 or #10 or #11 or #12
#14	#3 and #8 and #13
#15	MeSH descriptor: [Decision Making] explode all trees
#16	MeSH descriptor: [Decision Support Techniques] explode all trees
#17	"decision mak* OR shared decision making" or "medical decision making" or "choice behaviour" or "problem solving" or "clinical decision analysis" or "critical think*" or "decision aids" or "Task performance and analysis"
#18	#15 or #16 or #17
#19	#14 and #18

Appendix J. Cochrane Arbitration Search

Last Saved: 18/01/2016 19:02:17.644
Description:
ID Search
#1 MeSH descriptor: [Breast Neoplasms] explode all trees
#2 breast adj3 (neoplasm* or carcinoma* or tumour* or tumor* or cancer*)
#3 #1 or #2
#4 MeSH descriptor: [Mass Screening] explode all trees
#5 breast adj3 (scan* or screen* or radiograph* or imaging or visualise or visualize or exam* or test* or mammogra* or routine* or check* or diagnos* or detect*)
#6 mammogra* adj3 (scan* or screen* or visualise or visualize or exam* or test* or breast*)
#7 MeSH descriptor: [Mammography] explode all trees
#8 #4 or #5 or #6 or #7
#9 "Early detection of cancer*"
#10 "National Health Service Breast Screening Program" or "NHSBSP"
#11 "UK breast screen* program*"
#12 "NHS breast screen* program*"
#13 #9 or #10 or #11 or #12
#14 #3 and #8 and #13
#15 MeSH descriptor: [Negotiating] explode all trees
#16 arbitration* or arbitrat* or discordan* or discrepan* or disparity* or negotiat* or disagree* or conflict* or differen* or inconsisten* or variation* or consensus* or uncertain*
#17 #15 or #16
#18 #14 and #17

Appendix K. Cochrane Arbitration and Double Reading Search

Last Saved: 19/01/2016 16:24:52.576
Description:
ID Search
#1 MeSH descriptor: [Breast Neoplasms] explode all trees
#2 breast adj3 (neoplasm* or carcinoma* or tumour* or tumor* or cancer*)
#3 #1 or #2
#4 MeSH descriptor: [Mass Screening] explode all trees
#5 breast adj3 (scan* or screen* or radiograph* or imaging or visualise or visualize or exam* or test* or mammogra* or routine* or check* or diagnos* or detect*)
#6 mammogra* adj3 (scan* or screen* or visualise or visualize or exam* or test* or breast*)
#7 MeSH descriptor: [Mammography] explode all trees
#8 #4 or #5 or #6 or #7
#9 "Early detection of cancer*"
#10 "National Health Service Breast Screening Program" or "NHSBSP"
#11 "UK breast screen* program*"
#12 "NHS breast screen* program*"
#13 #9 or #10 or #11 or #12
#14 #3 and #8 and #13
#15 MeSH descriptor: [Negotiating] explode all trees
#16 arbitration* or arbitrat* or discordan* or discrepan* or disparity* or negotiat* or disagree* or conflict* or differen* or inconsisten* or variation* or consensus* or uncertain*
#17 #15 or #16
#18 MeSH descriptor: [Image Interpretation, Computer-Assisted] explode all trees
#19 "double-blind method*" or "double read*" or "double report*" or "reproducibility
of result*" or "sensitivity and specificity" or volume* or fatigue* or optimal
#20 #18 or #19
#21 #14 and #17 and #20

Appendix L. Scopus Search

3	TITLE-ABS-KEY (**mammography** AND **"decision making"** AND **reporting**)	02 Jun 2016	62
2	TITLE-ABS-KEY (**mammography** AND **"collective intelligence"**) Create a new term-based search | Set alert | Set feed	09 Mar 2016 last run on View new results	3 documents
1	(TITLE-ABS-KEY (**"mammography"**)) AND (TITLE-ABS-KEY (**"arbitration"**)) AND ((TITLE-ABS-KEY (**"breast cancer"**)) OR (TITLE-ABS-KEY (**"breast neoplasm"**)))	04 Jun 2016	36

Appendix M. Web Of Science Search

Set	Web of Science Search History - " arbitration 05/16"
#5	#3 AND #2 AND #1 Refined by: **LANGUAGES:** (ENGLISH) *DocType=All document types; Language=All languages;*
#4	#3 AND #2 AND #1 *DocType=All document types; Language=All languages;*
#3	**TOPIC:** (Negotiation or Negotiations or Mediation or Mediating or Arbitrating or Arbitration or Conflict Resolution or Conflict Resolutions or Resolution, Conflict) *DocType=All document types; Language=All languages;*
#2	**TOPIC:** (Mammographies or mammogram or mammography) *DocType=All document types; Language=All languages;*
#1	**TOPIC:** (breast neoplasm or breast carcinoma or breast tumour or breast cancer) *DocType=All document types; Language=All languages;*

Appendix N. CASP Diagnostic Checklist

Critical Appraisal Skills Programme

12 questions to help you make sense of a diagnostic test study

How to use this appraisal tool

Three broad issues need to be considered when appraising a diagnostic test:

- **Are the results of the study valid?** **(Section A)**
- **What are the results?** **(Section B)**
- **Will the results help me and my patients/population?** **(Section C)**

The 12 questions on the following pages are designed to help you think about these issues systematically.
The first two questions are screening questions and can be answered quickly. If the answer to both is "yes", it is worth proceeding with the remaining questions.

There is some degree of overlap between the questions, you are asked to record a "yes", "no" or "can't tell" to most of the questions. A number of italicised prompts are given after each question. These are designed to remind you why the question is important. Record your reasons for your answers in the spaces provided.

These checklists were designed to be used as educational tools as part of a workshop setting
There will not be time in the small groups to answer them all in detail!

The 12 questions are adapted from: Jaesche R, Guyatt GH, Sackett DL, Users' Guides to the Medical Literature, V1. How to use an article about a diagnostic test. JAMA 1994; 271 (5): 389-391

(A) Are the results of the study valid?

Screening Questions

1. Was there a clear question for the study to address? ❒Yes ❒Can't tell ❒No

HINT: A question should include information about
- The population
- The test
- The setting
- The outcomes

2. Was there a comparison with an appropriate reference standard? ❒Yes ❒Can't tell ❒No

HINT: Is this reference test(s) the best available indicator in the circumstances?

Is it worth continuing?

Detailed questions

3. Did all patients get the diagnostic test and reference standard? ❒Yes ❒Can't tell ❒No

HINT: Consider
- Were both received regardless of the results of the test of interest
- Check the 2X2 table (verification bias)

---

4. Could the results of the test have been influenced by the results of the reference standard? ❒Yes ❒Can't tell ❒No

HINT: Consider
- Was there blinding?
- Were the tests performed independently
- (Review bias)

---

5. Is the disease status of the tested population clearly described? ❒Yes ❒Can't tell ❒No

HINT: Consider
- Presenting symptoms
- Disease stage or severity
- Co-morbidity
- Differential diagnoses (Spectrum Bias)

6. Were the methods for performing the test Described in sufficient detail? ❒Yes ❒Can't tell ❒No

HINT: Consider
- Was a protocol followed?

(B) If so, what are the results?

7. What are the results?

HINT: Consider
- Are the sensitivity and specificity and/or likelihood ratios presented?
- Are the results presented in such a way that We can work them out?

---

8. How sure are we about the results? consequences and cost of alternatives performed?

HINT: Consider
- Could they have occurred by chance?
- Are there confidence limits?
- What are they?

(C) Will the results help me and my patients/population?

(Consider whether you are primarily interested in the impact on a population or individual level)

9. Can the results be applied to your patients/the population of interest? ❒Yes ❒Can't tell ❒No

HINT: Do you think your patients/population are so different from those in the study that the results cannot be applied? Such as age, sex, ethnicity and spectrum bias.

10. Can the test be applied to your patient or population of interest? ❒Yes ❒Can't tell ❒No

HINT: Consider

- Resources and opportunity costs
- Level and availability of expertise required to Interpret the tests
- Current practice and availability of services

11. Were all outcomes important to the individual or population considered? ❒Yes ❒Can't tell ❒No

HINT: Consider

- Will the knowledge of the test result improve patient wellbeing?
- Will the knowledge of the test result lead to a change in patient management?

12. What would be the impact of using this test on your patients/population?